VALIDATING THE CANADIAN ACADEMIC ENGLISH LANGUAGE ASSESSMENT FOR DIAGNOSTIC PURPOSES FROM THREE PERSPECTIVES: SCORING, TEACHING, AND LEARNING

by

Christine Diane Doe

A thesis submitted to the Faculty of Education in conformity with the requirements for the degree of Doctor of Philosophy

Queen’s University
Kingston, Ontario, Canada
(April, 2013)

Copyright ©Christine Doe, 2013
Abstract

Large-scale assessments are increasingly being used for more than one purpose, such as admissions, placement, and diagnostic decision-making, with each additional use requiring validation regardless of previous studies investigating other purposes. Despite this increased multiplicity of test use, there is limited validation research on adding diagnostic purposes—with the intention of directly benefiting teaching and learning—to existing large-scale assessments designed for high-stakes decision-making. A challenge with validating diagnostic purposes is to adequately balance investigations into the score interpretations and the intended beneficial consequences for teachers and students. The Assessment Use Argument (AUA) makes explicit these internal and consequential validity questions through a two-stage validation argument (Bachman & Palmer, 2010). This research adopted the AUA to examine the appropriateness of the Canadian Academic English Language (CAEL) Assessment for diagnostic purposes. To achieve this goal, I constructed a validity argument that answered the questions, to what extent did the CAEL essay meet the new diagnostic scoring challenges from the rater perspective, and to what degree were the teachers and students able to use the diagnostic information obtained from the CAEL. This study employed three research phases at an English for Academic Purposes (EAP) program in one Canadian university. Data collection strategies included interview and verbal protocol data from two raters (Phase 1), interview and classroom observation data from one EAP course instructor (Phase 2), and interview and open-ended survey data from 47 English Language Learners (Phase 3). A multifaceted perception of CAEL for diagnostic purposes was observed: raters noted the greatest diagnostic potential at higher score levels, and teacher and student perceptions were largely influenced by previous diagnostic assessment experiences. This
research emphasized the necessity of including multiple perspectives across contexts to form a deeper realization of the inferences and decisions made from diagnostic results.
Acknowledgements

First and foremost, I would like to acknowledge and thank Liying Cheng, my supervisor and mentor, for her guidance. Throughout this process she stimulated and stretched my thinking to consider new directions while at the same time supporting my choices and perspective. No matter how busy she was, Liying always made herself available to discuss and review my dissertation. I am grateful for her advice that was always tempered with reality on how to create a work-life balance in academia. I look forward to our future conversations and collaborations. I am thankful to have had her as my mentor and my gratitude cannot be captured in words alone.

I would also like to thank my committee members. Lyn Shulha continually pushed me to consider alternate explanations and directions with my work. John Freeman always found time to provide thoughtful and critical feedback on my dissertation or anything else I was working on. I am grateful to my examining committee, Michel Laurier, Sue Fostaty Young, Derek Berg, and Scott-Morgan Straker, for providing a stimulating conversation at my defence.

A number of individuals have supported me as a scholar over the years. Janna Fox introduced me to the world of language testing and inspired me to pursue my PhD. I could always count on Don Klinger for an engaging conversation that would challenge and push my thinking deeper. Rebecca Luce-Kapler supported my development as a teaching scholar for which I am truly grateful. On my first visit to Queen’s University, I met with Nancy Hutchison who welcomed me to the community with open arms. I would also like to thank my friends and colleagues at Mount Saint Vincent University for making me feel at home from day one.

I am grateful for the financial support I received to support my research and funding for the PhD. The CAEL office allowed me access to their test and provided support without which this research would not have been possible. I am also grateful for the 2008-2012 Ontario
Graduate Scholarships, the 2007-2008 Martin Schiralli Fellowship, and the 2011 Test of English as a Foreign Language small grant for Doctoral Research from the Educational Testing Service.

The Faculty of Education has a supportive graduate community, of which I was proud to be a member, that truly celebrates each other’s accomplishments and supports its members through any challenges. I am grateful to all its members for making my experience at Queen’s University memorable. There are a few individuals who I would like to thank by name. Michelle Searle is without a doubt one of the most magnanimous individuals I have ever met. Though we may no longer share the same address, I know our kindred spirits will remain just that for the rest of our lives. Her kindness and commitment to make the world a better place is larger than life. Alicia Hussain is someone I look to for strength and courage. I continue to treasure our long conversations about research and life. I am grateful to Christopher DeLuca for his honest feedback about my research ideas, presentations, and papers. His outlook and perspective is always refreshing.

My family has been a constant support throughout my life. I am grateful to my grandparents, Mothers and Pas, for encouraging me in all my endeavours— from building a tree house to travelling the world; to my aunt and uncle, Elaine and Dennis, for helping me celebrate my successes along the way; to my uncle, Don, for being an inspiration on how to advocate for others; to my cousins, Neil and Doug, for being like brothers to me; to my stepfather, Dave, for supporting me in every way possible; to my sisters, Sacha and Afton, two amazing women I am proud to be related to; and to my dad and stepmother, Keath and Debbie, for encouraging me to pursue graduate work.

And finally, to my mother, Diane—my sounding board, champion, and friend—she has always believed in me. I see the world as my playground because of her.
# TABLE OF CONTENTS

Abstract ................................................................................................................................. i
Acknowledgements ............................................................................................................... iii
Table of Contents ..................................................................................................................... v
List of Tables ........................................................................................................................... ix
List of Figures ........................................................................................................................ x

## Chapter 1: Introduction ........................................................................................................ 1
  - Background ....................................................................................................................... 1
  - Purpose .............................................................................................................................. 2
  - Research Questions ......................................................................................................... 3
  - Rationale .......................................................................................................................... 3
  - Defining Diagnostic Assessment ...................................................................................... 4
    - Content ........................................................................................................................ 5
    - Grain-size ....................................................................................................................... 6
    - Delivery ......................................................................................................................... 6
    - Intended use ................................................................................................................... 7

## The Research Context ........................................................................................................ 7
  - The Canadian Academic English Language (CAEL) assessment ..................................... 7
  - The Canadian Academic English Language Diagnostic (CAELD) assessment .............. 8
  - The English for Academic Purposes (EAP) program ..................................................... 10

## Chapter 2: Theoretical and Empirical Rationale ............................................................... 12
  - Validity Theory and Frameworks .................................................................................... 12
    - Test use ........................................................................................................................ 13
    - Consequences from test use ....................................................................................... 14
    - Constructing a case for test use and beneficial consequences .................................... 17
  - Review of Research Literature ....................................................................................... 22
    - Scoring ........................................................................................................................ 22
    - Teaching ....................................................................................................................... 28
    - Learning ....................................................................................................................... 36
  - The Assessment Use Argument for the CAELD .............................................................. 38
    - Beneficial consequences ............................................................................................. 39
    - Decisions ...................................................................................................................... 39
    - Interpretations .............................................................................................................. 40
    - Assessment Records .................................................................................................... 41

## Summary ........................................................................................................................... 42

## Chapter 3: Methodology .................................................................................................... 43
  - Qualitative Case Approach ............................................................................................ 43
  - Participants ..................................................................................................................... 46
    - Raters ........................................................................................................................... 46
    - Teacher ........................................................................................................................ 46
    - Students ....................................................................................................................... 48
  - Instruments ..................................................................................................................... 49
    - Phase 1 ......................................................................................................................... 49
    - Phase 2 ......................................................................................................................... 52
Chapter 4: Scoring

Data Collection
Phase 1 ................................................................................................................. 55
Phase 2 ................................................................................................................ 57
Phase 3 ................................................................................................................ 59

Data Analysis
First cycle analysis ............................................................................................. 61
Second cycle analysis ......................................................................................... 62
Third cycle analysis ........................................................................................... 68

Criteria for Judging Qualitative Evidence ......................................................... 69

Chapter 4: Scoring

Summary ............................................................................................................. 70

Chapter 5: Teaching

Overview ........................................................................................................... 98

Research Question ........................................................................................... 99
Themes and Codes .............................................................................................. 99

Background and Context
Approach to teaching and learning ................................................................ 101
Approach to assessment ..................................................................................... 103
Teaching context and course goals ................................................................. 104

The Decision-Making Process
Planning ............................................................................................................ 105
Evaluation ......................................................................................................... 106
Skills and strategies .......................................................................................... 107
Instructional support ......................................................................................... 111
Using the CAELD ............................................................................................. 114
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Examining the CAELD</td>
<td>190</td>
</tr>
<tr>
<td>Chapter 8: Conclusion and Implications</td>
<td>193</td>
</tr>
<tr>
<td>Beneficial Consequences in situ</td>
<td>194</td>
</tr>
<tr>
<td>Validity Argument(s)</td>
<td>198</td>
</tr>
<tr>
<td>A way forward</td>
<td>201</td>
</tr>
<tr>
<td>References</td>
<td>203</td>
</tr>
<tr>
<td>Appendices</td>
<td></td>
</tr>
<tr>
<td>Appendix A: CAEL Score Report</td>
<td>220</td>
</tr>
<tr>
<td>Appendix B: Summary of Skill Subcomponents Augmented by Literature</td>
<td>221</td>
</tr>
<tr>
<td>Appendix C: Reading, Listening, and Writing Skill Subcomponents</td>
<td>223</td>
</tr>
<tr>
<td>Appendix D: Learner Profile (Teacher Version)</td>
<td>225</td>
</tr>
<tr>
<td>Appendix E: Learner Profile (Student Version)</td>
<td>227</td>
</tr>
<tr>
<td>Appendix F: Assessment Use Argument for the CAELD</td>
<td>229</td>
</tr>
<tr>
<td>Appendix G: Ethics Clearance and CHRPP Certificate</td>
<td>232</td>
</tr>
<tr>
<td>Appendix H: Instructions for Reviewing Essays (Phase 1)</td>
<td>235</td>
</tr>
<tr>
<td>Appendix I: Rater Interview Protocol (Phase 1)</td>
<td>236</td>
</tr>
<tr>
<td>Appendix J: Instructor Interview Protocol (Phase 2)</td>
<td>238</td>
</tr>
<tr>
<td>Appendix K: Classroom Observation Protocol (Phase 2)</td>
<td>244</td>
</tr>
<tr>
<td>Appendix L: Student Reflections (Phase 3)</td>
<td>245</td>
</tr>
<tr>
<td>Appendix M: Students’ Interview Protocol (Phase 3)</td>
<td>251</td>
</tr>
<tr>
<td>Appendix N: Features of Writing Subcodes</td>
<td>256</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1: Placement into the EAP Program based on CAEL Results ...........................................11
Table 2: Facets of Validity from Messick (1989, p. 20) .................................................................15
Table 3: Variations of Formative Assessment Characteristics (McMillan, 2010, p. 43) 29
Table 4: Demographic Information of Students Interviewed .........................................................49
Table 5: Research and Facilitating Questions and Corresponding Data Sources .......................51
Table 6: Essays Included for the Think Aloud Protocols ............................................................56
Table 7: Summary of Phase 2 Data Collected .............................................................................58
Table 8: Phase 1 Coding Framework .........................................................................................63
Table 9: Phase 2 First and Second Cycle Coding Framework ....................................................66
Table 10: Phase 3 Coding Framework .........................................................................................68
Table 11: Frequency Counts of Phase 1 Themes and Codes .........................................................74
Table 12: Difference Between Original and Think Aloud (New) Score .....................................76
Table 13: Features of Writing Ability Discussed by the Raters During the Think Alouds .......77
Table 14: Frequency of Dimensions of Student Writing Ability by Level of Essay .................83
Table 15: Types of Advice Offered by the Raters .......................................................................88
Table 16: Overview of Advice Type by Level of Essay ...............................................................90
Table 17: Phase 2 Themes and Codes with Frequencies ..........................................................100
Table 18: Diagnostic Subcomponents Targeted ........................................................................110
Table 19: Phase 3 Themes and Codes with Frequencies ...........................................................134
Table 20: Students’ Rank Order of Diagnostic Instructional Activities ..................................143
Table 21: Students’ Perceived Strengths (13 total) .................................................................151
Table 22: Students’ Perceived Weaknesses (54 total) ..............................................................152
Table 23: Students’ Perceptions of Diagnostic Assessment at the Beginning and End of the Course .............................................................................................................155
Table 24: Approaches Students Identified Helping Them Target Their Weaknesses (Review and Learn; Practice) ..................................................................................158
Table 25: Approaches Students Identified Helping Them Target Their Weaknesses (Strategy; Engage in Further Learning) .................................................................159
Table 26: Assessment Claims for the AUA for the CAELD .........................................................165
Table 27: Assessment Records Warrant (D.1) for Assessment Claim D ..................................166
Table 28: Meaningfulness Warrants (C.1–C.2) for Assessment Claim C ................................168
Table 29: Meaningfulness Warrants (C.3–C.4) for Assessment Claim C ................................170
Table 30: Meaningfulness Warrants (C.5–C.6) for Assessment Claim C ................................173
Table 31: Impartial Warrants (C.7) for Assessment Claim C .....................................................174
Table 32: Impartial Warrants (C.8–C.11) for Assessment Claim C ...........................................175
Table 33: Generalization Warrants (C.12–C.13) for Assessment Claim C ................................177
Table 34: Relevance Warrants (C.14–C.15) for Assessment Claim C .........................................178
Table 35: Relevance Warrant (C.16) for Assessment Claim C ..................................................179
Table 36: Sufficiency Warrant (C.17) for Assessment Claim C ................................................180
Table 37: Decisions Warrant (B.1) for Assessment Claim B .....................................................181
Table 38: Decisions Warrants (B.2–B.4) for Assessment Claim B ............................................183
Table 39: Beneficial Consequences Warrants (A.1–A.3) for Assessment Claim A ................185
Table 40: Beneficial Consequences Warrant (A.4) for Assessment Claim A ..........................188
Table 41: Beneficial Consequences Warrant (A.5) for Assessment Claim A ..........................189
LIST OF FIGURES

Figure 1: Construction and Evaluation of an Assessment Use Argument..........................20
Figure 2: Example Assessment Use Argument for the CAELD.................................21
Figure 3: Reflection 1 student response counts for if they thought the CAELD feedback reflected their actual language ability..........................................................149
Figure 4: Reflection 2 and 3 student responses indicating if they thought about the CAELD feedback during the Intermediate-level EAP course.........................153
Figure 5: Reflection 2 and 3 student responses indicating if they thought about the CAELD feedback on their own and outside of class.................................154
Chapter 1: Introduction

Background

My fascination with this research topic and the broader field of language testing started as an English language teacher. While on a plane for 13 hours, en route to South Korea, I could not sleep. I was nervous about whether or not I would be a good teacher. Instead, I read newly purchased books on how to teach English. I soon realized that, while the books presented me with fun activities, they glossed over some of the challenges inherent in teaching: they offered little guidance on pinpointing the types of errors my students might make, or how I could help my students improve.

As I continued teaching in Korea, I began to realize that the skills necessary to be an effective language teacher were fundamentally about communication. While I was able to teach the mechanics of what was needed well enough, I struggled to convey the nuances my students needed to truly be successful users of English. I observed my students employ a range of strategies to try to improve. I was particularly amazed by one student, Young ho, and his tactic for passing the Test of English for International Communication (TOEIC). He would go to the windowless, basement listening-lab six days a week to memorize recorded lectures, songs, and other forms of listening material. In the year that I knew Young ho, I watched him improve his language skills, successfully pass the TOEIC, and get a job with an international company in Seoul. Despite his successes, I could not help but wonder if there was an easier way.

On a quest to find a better way to help second language (L2) students learn English, I pursued my Master’s of Applied Language Studies at Carleton University. One of the assignments for my final course in the program was to research a well-established language-testing expert. I chose Elana Shohamy as I was particularly drawn to her 2001 article,
Democratic Assessment as an Alternative, in which she argued that the stakeholder voice was being underutilized in test development and validation studies. I later heard Dr. Shohamy speak at the 2005 Language Testing Research Colloquium in Ottawa with her presentation on “Tests as Power Tools: Looking Backwards, Looking Forward.” Her passion and enthusiasm for the topic were astounding. She highlighted the numerous misuses of assessments and the negative impact they have had on language learners, offering No Child Left Behind and citizenship tests in the United States as examples. In the face of this profound negativity surrounding language testing, Shohamy (2005) concluded the presentation with a call to action. She reiterated the need to include key stakeholders (e.g., teachers and students) in test development and validation work, but also spoke about the capacity that language assessments have to promote positive change (Shohamy, 2001).

Inspired by Shohamy’s (2005) call for action, my master’s thesis investigated students’ learning experiences in preparing to take the Canadian Academic English Language (CAEL) assessment. Specifically, I looked at their strategy use across three testing contexts: a test preparation course, a practice test, and a live test (Doe & Fox, 2011). I was intrigued that one of the students did not take the CAEL with the hope of obtaining the score necessary for either admission into the English for Academic Purposes (EAP) program or full admission into the university; rather, this student took the CAEL to learn from it. She wanted to know how she needed to improve. This idea of learning from the test prompted my current interest in examining the potential of language assessments for diagnostic purposes.

Purpose

The purpose of this study is to examine the appropriateness of the Canadian Academic English Language Diagnostic (CAELD) assessment for diagnostic purposes in an English for
Academic Purposes (EAP) program. This study builds on my master’s thesis by examining the use of a large-scale English language proficiency assessment for diagnostic purposes. Drawing on an argument-based approach to validation (Bachman & Palmer, 2010), this study investigated the use of a large-scale language proficiency assessment, the Canadian Academic English Language Diagnostic (CAELD) assessment used for diagnostic purposes in an EAP program. The CAELD is the retrofitted version of the Canadian Academic English Language (CAEL) assessment—a large-scale assessment used for admissions and placement purposes in an EAP program (Fox, 2009). The study explored the decisions related to scoring, teaching, and learning and the extent to which these processes contributed to teaching and learning.

**Research Questions**

1. What are the features of CAELD scoring that enable it to be used for diagnostic purposes?
2. What are the teaching conditions that enable the CAELD results to be used for diagnostic purposes?
3. What are the learning conditions and students’ perceptions that enable the students to use the CAELD and its results to support their learning?

 Each of these broad research questions guides a phase of this study. Each question is addressed through specific facilitating questions as outlined in Chapter 3 and discussed in Chapters 4, 5, and 6.

**Rationale**

The number of L2 students studying at English-speaking universities around the world is continually increasing. From 2005 to 2010, the number of international students studying at post-secondary institutions in Canada increased by more than 45,000 (Statistics Canada, 2013). Often
not counted within this L2 population are the immigrant students who also speak English as a second language, but completed some of their K-12 education in Canada (Abada, Hou, & Ram, 2008; McMullen & Elias, 2011). Indeed, 79% of immigrant youth in Canadian high schools plan on attending university (Statistics Canada, 2007; Taylor & Krahn, 2005). Both groups add to the increasing number of L2 students on Canadian university campuses. All L2 students intent on full-time study at an English-speaking, Canadian university have to satisfy a language requirement with either (a) proof of three or more years attendance in an English-speaking high school, (b) a passing score on a language proficiency assessment, such as, the CAEL, or (c) in the case of this study’s university context, successfully completion of the EAP program. L2 students who have not yet satisfied the requirement are generally highly motivated to improve their English as quickly as possible, in order to satisfy the language requirement through option (b) or (c).

By retrofitting the CAEL to have diagnostic properties (CAELD), students may have the opportunity to receive detailed feedback about their language performance on the assessment, which they may use to refine their studying and, ultimately, improve their English skills. Further, the diagnostic feedback, combined with targeted instruction that focuses on students’ specific academic language weaknesses, would be expected to expedite the language learning process. Increasing evidence demonstrates the potential of drawing on such diagnostic feedback to inform teaching and student learning (e.g., Fox, 2009; Jang, 2009).

**Defining Diagnostic Assessment**

Defined simply, *diagnostic assessment* refers to the process of identifying students’ linguistic strengths and weaknesses (Alderson, 2005). It is possible to diagnose students’ linguistic strengths and weaknesses through a variety of assessments. Indeed, the concept of
diagnostic assessment has often been compared to assessments used for placement or formative purposes (see Alderson, 2005; Huhta, 2008). Both purposes assess gaps in learning to ultimately establish learning opportunities appropriate for students’ proficiency levels and/or learning needs. In the case of placement testing, students are typically grouped into courses based on similar proficiency levels; it is assumed that the students would have similar learning needs, so the instruction should be suitable for all students. Within the classroom, teachers would draw on formative assessment techniques to identify possible gaps between course instruction and the students’ learning needs (Black & Wiliam, 1998; Huhta, 2008). Thus both of these assessment scenarios involve identifying the students’ learning needs, through diagnosing their abilities, though the level of specificity about students’ strengths and weaknesses varies.

While more and more assessments are being designed and constructed for the sole objective of diagnosing student language ability (e.g., Alderson, 2005; Read, 2008), there are also numerous assessments that take on diagnostic properties in a variety of ways (Kunnan & Jang, 2009). There are four features of a language assessment, such as the CAELD, that are considered when designing or adapting a test for diagnostic purposes: content, grain-size, delivery, and intended use.

- **Content.** The assessment may be based on either a general model of language proficiency or on a curriculum (Jamieson, Grgurovic, & Becker, 2008). For instance, DIALANG, a free online assessment used to diagnose 14 languages, is based on can-do statements from the Common European Framework of Reference (CEFR) (see Alderson, 2005). In contrast, the NorthStar Adaptivity project offers an example of a curriculum-based assessment (Jamieson, 2011; Jamieson et al.). With the assessment based on the NorthStar textbooks for L2 students, the identification of students’
strengths and weaknesses from the diagnostic assessment is directly connected to the curriculum taught.

- **Grain-size.** Grain-size refers to the level of detail given in the feedback, which may be coarse or fine (Jang, 2008). If the feedback is by skill level, then it is coarse, similar to an assessment used for placement purposes (Bachman, 1990); if the feedback is by skill subcomponent, it is fine-grained. Another aspect of grain-size is how strengths and weaknesses are defined and determined. Recent advancements in psychometrics have resulted in sophisticated techniques to identify fine-grained diagnostic skills through Q-matrices using the cognitive diagnostic assessment framework (Jang, 2008). Jang (2005, 2009) used the terms masters and non-masters to describe students who were strong or weak in particular subcomponents. Drawing on cognitive diagnostic assessment theory, Jang (2005) was able to identify masters and non-masters of reading questions based on a probability statistic with three cut-off points: master (>0.6), undetermined (0.4–0.6), and non-master (<0.4). In contrast, Poehner (2008) describes a more subjective approach to identifying students’ strengths and weaknesses using dynamic assessment. The authors acknowledge the concern for lower reliability across classroom settings, but argue that the on-the-spot nature of dynamic questioning allows the assessor to better diagnose the students’ true ability.

- **Delivery.** Diagnostic feedback is often given to learners (and/or teachers) in the form of an online or paper-based learning profile, which includes descriptions of the learner’s strengths and weaknesses, listed by skill and subcomponent (e.g., Alderson, 2005; Fox, 2009; Jang, 2009; Read, 2008). The feedback given to the teacher and to
the student may differ in format, in order to best facilitate understanding. Diagnostic feedback may also be given in-person through dynamic assessment, where the evaluator, most likely the teacher, is the diagnosing instrument rather than a paper-and-pencil test (Leung, 2007).

- **Intended use.** Diagnostic feedback may be used in a variety of ways. For instance, Shohamy (1992) and Fox (2009) look to language diagnosis as a pivot point for program evaluation or curricular renewal. Read (2008) and colleagues describe the specialized language support students receive, typically through courses, based on the two-phase Diagnostic English Language Needs Assessment (DELNA) program. Numerous researchers describe in-class formative purposes as the commonly desired use of diagnostic feedback (Alderson, 2007; Jamieson et al., 2008; Jang, 2009).

By drawing on CAEL raters’ expertise in scoring CAEL essays and teaching, this study explored the connection between the scoring of the CAEL essays for diagnostic purposes and the possible advice offered to students on how to improve. In addition, the CAELD should support teachers and students in their teaching and studying of English.

**The Research Context**

The research context included the CAEL assessment, the CAELD, and the EAP program where the study took place.

**The Canadian Academic English Language (CAEL) assessment.** The CAEL, a large-scale language proficiency assessment, measures English use in academic settings for university admissions and EAP placement purposes (CAEL Assessment Office, 2008). The assessment was designed to elicit academic English language skills from test takers in two ways. First, the assessment simulates an academic setting, in which test takers complete various tasks related to
reading, listening, and speaking. These tasks mirror those university students are required to perform on a daily basis (e.g., taking notes while listening to a lecture). Second, test takers must integrate ideas from the reading and listening sections to formulate an argumentative essay. Thus the CAEL is considered a task- and topic-based, fully integrated assessment.

The CAEL is composed of the Written Test and the Oral Language Test. The Written Test consists of three sections on reading, listening, and writing; included in the respective sections are two readings of 50 minutes, one listening that is a lecture of about 18 to 19 minutes in length, and a written essay of 45 minutes (CAEL Assessment Office, 2008). The reading, listening, and writing are based on one topic. For the written essay, test takers are required to write an argumentative essay based on the readings and listening. The prompt for the essay is given at the beginning of the Written Test, so the test takers can think about what information they may use from the listening and reading to write their essays. Scoring of the CAEL is represented by nine band scores ranging from 10 to 90 (see Appendix A). A total band score is reported, as well as four band scores for each of the subtests (i.e., reading, listening, writing, and speaking). Raters arrive at a writing band score using a scoring guide divided into the categories of language, organization, and content. The Oral Language Test and the Written Test are administered on separate days, and are based on separate topics.

The Canadian Academic English Language Diagnostic (CAELD) assessment. In 2009, the CAEL was retrofitted for diagnostic purposes to promote curricular renewal in an EAP program (see Fox, 2009). The name CAELD was adopted to distinguish this test from the CAEL, which is used for placement or admissions purposes. To retrofit the CAEL, Fox (2009) drew on the pilot test specifications to indicate the tasks/questions corresponding to particular skill subcomponents (e.g., reading for the main idea). The CAEL test specifications document
outlines the structure of the test and the skills that each question is designed to elicit (Alderson, Clapham, & Wall, 2003). In addition to the test specifications, Fox (2009) drew on research investigating the underlying skill subcomponents, namely, by Abbott, 2007; Banerjee and Wall, 2006; Buck and Tatsuoka, 1998; Jang, 2005 (see Appendix B for the research studies supporting the subcomponents and Appendix C for the underlying CAELD subcomponents for reading, listening, and writing). Due to the challenges of providing diagnostic feedback on speaking and the limited research to connect the test specifications to skill subcomponents with confidence, the speaking section was not included.

For this study, learning profiles were created for teachers to use diagnostically in the teachers’ classrooms following the procedures outlined by Fox (2009). In addition, a student version of the learner profile was created to provide students with feedback to aid in their study of academic English (see Appendices D and E for example learner profiles for the teacher and students). Prior to formalizing the research design, I met with the teacher. We discussed how she would like to have the diagnostic feedback represented in her class. She wanted to narrow the number of skill subcomponents reported (from 10 to 4 for reading, and from 10 to 8 for listening). Involving the teacher in the development and creation of the learner profiles set this study apart from other studies that use large-scale tests as a diagnostic assessment.

At the time of the study, there had been two previous occasions when the CAEL was used as a diagnostic assessment to inform classroom instruction (Fox, 2009; J. D. Fox, personal communication, December 4, 2009). The first time was in response to policy changes regarding the placement of students into an EAP program. In conjunction with a validity study of the placement decisions, Fox (2009) retrofitted the CAEL to develop diagnostic profiles for all students in the program. The second time occurred on a smaller scale, with one teacher using the
diagnostic feedback to inform the teaching of two EAP classes. The study here was based on these two previous instances with its purpose to examine the appropriateness of the CAELD for diagnostic purposes in an EAP program, and, in turn, to develop a deeper understanding of large-scale testing for diagnostic uses by looking at the aspects of scoring, teaching, and student learning.

**The English for Academic Purposes (EAP) program.** This study took place in the EAP program offered at a Canadian university in south-eastern Ontario. The CAEL is one of a number of possible tests that place students into this EAP program. The CAEL office is housed in the same building as the EAP program. This proximity allows for collaboration between the CAEL testing program and the EAP program.

The English for Academic Purposes program consists of three levels of courses (beginner, intermediate, advanced). Entrance into the three levels is based on test scores obtained from one of the recommended language proficiency tests, for example, Test of English as a Foreign Language (TOEFL), International English Language Testing System (IELTS), and CAEL. If a student obtains a high enough score on the CAEL, he or she will be admitted into full-time academic study and will not be required to take any EAP courses. Students entering undergraduate programs at the university in my study can gain full admission into their program by successfully completing the advanced level EAP course or by receiving a CAEL score of 70. For students without the requisite CAEL scores to gain full admission, admittance into the EAP program offers a highly desirable alternative, as they are able to begin taking credit academic courses alongside their EAP courses; this process is considered to be graduated admission into the university. In addition, the university in my study will accept the EAP courses for credit
towards degrees in some programs. Placement of students into the three EAP courses, as determined by CAEL band score results, is summarized in Table 1.

Table 1

*Placement into the EAP program based on CAEL results*

<table>
<thead>
<tr>
<th>EAP course level</th>
<th>Overall CAEL Band score</th>
<th>Number of undergraduate courses EAP students are permitted to take</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beginning</td>
<td>30, 40</td>
<td>1</td>
</tr>
<tr>
<td>Intermediate</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td>Advanced</td>
<td>60</td>
<td>3</td>
</tr>
<tr>
<td>Full admission into</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>undergraduate/graduate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>programs</td>
<td></td>
<td>Full-time status</td>
</tr>
</tbody>
</table>
Chapter 2: Theoretical and Empirical Rationale

This chapter provides the theoretical and empirical grounding for the study. To contextualize the theoretical orientation and framework taken, I first provide an overview of the key theoretical and historical shifts in validity theory and validation. Second, I outline the research literature as it relates to using a learning-orientated assessment to inform scoring and teaching, thus supporting students in their acquisition of academic English. I conclude the chapter by outlining the specific validity framework, the Assessment Use Argument (AUA), that I adopted for this study to examine the appropriateness of the CAELD for diagnostic purposes in an EAP program.

Validity Theory and Frameworks

Validity theory has evolved significantly from the 1950s to the present day, from a focus on primarily score-based predictions to integrated evaluative judgments that support test use (see Kane, 2006; Xi, 2008). Validity is the cornerstone of any testing program. Messick (1989) defined validity as an “integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment” (p. 13). Validation is the process of justifying the claims made about the test and its use (Bachman & Palmer, 2010). In this study, the validation of the CAELD as a diagnostic instrument depended on the extent to which the meaning of test-takers’ performances could be accurately interpreted and the extent to which an argument could be made that these interpretations were able to support teaching and learning in ways that contributed to improved language performance.

Three shifts in the theoretical conceptualization and application of validity theory have informed this study. First, the idea of validating an assessment as the unit of analysis has shifted
focus to the interpretations made about test takers’ performances for a specified purpose, in a particular context (Cronbach, 1971); in other words, the implications of validation research cannot extend to additional test uses not included in the design. Second, validity theory is now construed as a wider-reaching concept, one that looks beyond test takers’ performances to consider the social value given to tests and their impact on stakeholders—in the case of this study, the teacher and students (Messick, 1989). Third, in response to the need for validation research to be theoretical and empirically grounded, argument-based approaches were introduced as a way to connect the empirical research as chain of evidence in support of the theoretical interpretive claims made about an assessment (Kane, 2006).

**Test use.** The idea of validating test use, and not the test itself, emerged in Cronbach and Meehl’s (1955) introduction of construct validity to the testing community. Constructs describe an underlying trait that cannot be directly measured and must be operationalized to be measured through inferences (Fulcher & Davidson, 2007). Validation research up to this point was calculated through validity scores that correlated the test results with an external criterion in the form of future performance (predictive validity) or another test score (concurrent validity; Cureton, 1951; Moss, Girard, & Haniford, 2006). To establish content validity, test items were examined for their representativeness of the target domain (American Educational Research Association, 1999). Cronbach and Meehl (1955) emphasized the need to articulate the construct being measured by elaborating the types of inferences being made about test takers’ performances. They further stressed that, through this detailed specification of constructs, test developers were not validating “a test, but only a principle for making inferences” (p. 297).

In the second edition of *Educational Measurement*, Cronbach (1971) reaffirmed the notion of construct validity by encouraging researchers to refocus the purpose of validation
research towards the “educational uses and interpretations of tests” (p. 445). He was critical of the testing community for its continued, narrow focus on validation research as score-based predictions grounded in the concepts of criterion-reference and content validity. After Cronbach’s (1971) chapter, construct validity became the focal point for test development and validation research, and test developers became more specific about the purposes of an assessment. Now it is commonly understood that language assessments are not viewed and validated according to type, such as placement or admissions, but according to whether or not the interpretations about test taker performance are justified for the intended use (Xi, 2008). A tension arises when an assessment is used for a purpose not included in the original design and development, or a change is made to the test without adjusting the interpretative claims.

**Consequences from test use.** Messick (1989) in the third edition of *Educational Measurement* brought this social component to the forefront of validity-based inquiry. His *Facets of Validity* (see Table 2) outlined a four-cell validity matrix. Messick argued that validity needed to be interpreted holistically; for him it was necessary to gather evidence that filled the four cells to arrive at an integrated evaluation of the validity supporting test use. He acknowledged that, in reality, the validity evidence collected could not be spliced as cleanly as the table suggested; the complex social setting of the test and its outcomes would make the boundaries of the cells difficult to discern.

By including the social component within the same table as other subsets of validity, Messick (1989) brought the idea of value-based interpretations and consequential validity to the forefront of testing research, in educational measurement and language testing. In response to
Messick’s (1989) framework, language-testing researchers have expanded this social dimension of validity through key advancements in the language testing field: test usefulness, washback and impact research, and formative assessment.

Table 2

*Facets of Validity from Messick (1989, p. 20)*

<table>
<thead>
<tr>
<th>Evidential Basis</th>
<th>Test Interpretation</th>
<th>Test Use</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Construct Validity</td>
<td>Construct Validity +</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Relevance/Utility</td>
</tr>
<tr>
<td>Consequential Basis</td>
<td>Value Implications</td>
<td>Social Consequences</td>
</tr>
</tbody>
</table>

Bachman and Palmer (1996) proposed their test usefulness framework as a way to evaluate language tests across six qualities: reliability, construct validity, authenticity, interactiveness, impact, and practicality. In contrast to Messick’s (1989) validity framework, Bachman and Palmer (1996) categorized validity as being only related to construct validity. The social consequences piece was considered through the impact quality of test usefulness. Xi (2008) stated that the test usefulness model had served as the primary test development and validation framework in language testing for nearly two decades.

The social consequences cell of Messick’s (1989) facets of validity expanded into washback and impact research—a rich and thriving field of study unto itself in language testing—perhaps more than it expanded into educational measurement. Washback refers to the influence that tests have on teaching and learning (Alderson & Wall, 1993). McNamara and Roever (2006) attributed this divergence in research agendas for the two disciplines (language assessment and general education assessment) to the socially embedded nature of language:

Although the social dimension of assessment might not be such a pressing issue in the measurement of other areas, … the fact that language is so inextricably part of the fabric of, indeed constitutive of, social and cultural life, and of face-to-face interaction has
meant that the paradigmatic one-sidedness of conventional approaches to assessment comes into particularly sharp focus in language testing. (p. 247)

In this vein, impact and washback research has proved essential in uncovering some of the connections between high-stakes language testing, teaching, and learning, as situated within the broader social and political milieu (see Cheng, 2008; Cheng, Watanabe, & Curtis, 2004). Sometimes education systems will introduce a high-stakes test with the aim of promoting positive washback for teaching and learning (Qi, 2005). However, influencing teaching and learning through testing is often based on a number of other factors, such as teachers’ beliefs (Shohamy, Donitz-Schmidt, & Ferman, 1996). More recently, the concept of utilization or beneficence has been recognized as a key component of test validation—one that ensures that the consequences resulting from language tests are positive for test takers, without accompanying adverse effects (Bachman, 2005; Kunnan, 2004).

An outcome of the extensive washback and impact literature is that the importance of stakeholder perceptions and diverse methodologies was recognized. Indeed, data gathered from multiple stakeholders have resulted in deeper, and sometimes conflicting, explanations of test use (Qi, 2005). Qualitative and mixed methodologies sensitive to the dynamic, social language-testing contexts have been vital for revealing the complex stakeholder relationships that ultimately manifest test use (Cheng, 2008). The social dimension of validity has been a vital component in uncovering the complex and dynamic educational systems within which language tests operate (Cheng, 2008; Saville, 2010).

In the same year Messick wrote his chapter on validity, Sadler (1989) introduced the idea of formative assessment to the teaching and assessment community; this idea was later extended by Black and Wiliam (1998), with their assessment for learning theory. Now assessments are
framed as a fundamental component of learning cultures, with positive assessment experiences occurring prior to, throughout, and after instruction (Black, Harrison, Lee, Marshall, & Wiliam, 2003; Shepard, 2000). However, applying traditional notions of validity theory to dynamic assessment contexts is not without contention (Teasdale & Leung, 2000). Researchers have challenged the practice of applying psychometric traditions to classroom assessment processes, since this application may undermine teachers’ decision-making processes (Brookhart, 2003; Moss, 1996). The value and role of a large-scale assessment, such as the CAELD used for diagnostic purposes, would be influenced by the values attributed to the assessment and the process by the teacher and students.

Messick (1989) helped situate testing as an inherently social endeavour by expanding validity theory to consider the value implications and consequences of test use. A troubling outcome of this theoretical expansion was that it required an indeterminate amount of evidence for test validation research but provided little guidance on how to prioritize a program of research that linked to the inferences and decisions being made about test takers.

**Constructing a case for test use and beneficial consequences.** Argument-based approaches to validation provide a conceptual framework on how to construct a case that would convince test users that the claims about the intended uses are warranted. Key to developing an assessment argument is prioritizing the evidence needed to establish a well-supported case of test use. Argument-based approaches are quickly becoming the prevailing method for present-day validity inquiry in language testing (e.g., Bachman, 2005; Chapelle, Enright, & Jamieson, 2007; Colby-Kelly & Turner, 2007; Jang, 2005; Kane, 2012). These logical frameworks have roots in Cronbach and Meehl’s (1955), Cronbach’s (1971, 1984, 1989), and Messick’s (1989) work. Cronbach and Meehl (1955) emphasized the obligation for test developers to present the test user
with a summary describing the interpretations that are well supported alongside the alternative explanations, and those that are speculation. Cronbach (1971) advocated for validation research to draw on multiple sources of evidence in order to adequately support the interpretations made about test takers’ performances. Messick (1989) viewed test validation as an inquiry-based process that could draw on any number of theories “to marshall evidence and arguments in support of, or counter to, proposed interpretations and uses of test scores” (p. 32). In his chapter, Messick (1989) outlined five possible philosophical orientations that could guide the inquiry process, such as, Leibnizian or Hegelian inquiring systems. Missing from these iconic works on validity theory was explicit guidance on how to connect the claims about test use as well as how the interpretations supporting that use mapped onto the research evidence needed to justify the conclusions.

Educational researchers and language testers responded to this need to justify test use through validation research by introducing argument-based approaches that not only were grounded theoretically, but also offered logical steps for outlining a methodology to support the interpretations made about test taker performance (Kane, 1992, 2002, 2006). For this study, I adopted the Assessment Use Argument (AUA) proposed by Bachman and Palmer (2010). The AUA draws on Toulmin’s (2003) argumentative theories of logic, which are similar to Kane’s (2006) validity argument, Mislevy, Steinberg, and Almond’s (2003) evidence-centred design, and the validity argument for the TOEFL iBT (Chapelle et al., 2007). The AUA was chosen because it begins by considering the beneficial consequences for stakeholders as the first claim in the test development and validation process. This prominence given to beneficence (Bachman & Palmer, 2010; Kunnan, 2004) in AUA is in line with an effect-driven design (Fulcher &
Davidson, 2007), and with this study’s goal to examine the appropriateness of the CAELD for diagnostic purposes in an EAP program.

Establishing an AUA for test use is a two-stage process, in which the specific assessment claims are identified before evidence (backing) is gathered to support the underlying warrants on which the claims are based. Keeping with an effect-driven design and the recommendations from Bachman and Palmer (2010), the construction of the AUA for the CAELD begins with the intended beneficial consequences, while the evaluation of the argument follows the assessment process and begins with the assessment records. Figure 1 depicts this dual directionality of designing and then appraising an AUA based on the data collected. The letters assigned to the assessment claims are different from Bachman and Palmer (2010), since they attributed numbers to the claims.

Bachman and Palmer (2010) further detailed the four general claims as:

Claim A. The consequences of using an assessment and of the decisions that are made are beneficial to stakeholders.

Claim B. The decisions that are made on the basis of the interpretations

- take into consideration community values and relevant legal requirements and
- are equitable for those stakeholders who are affected by the decisions.

Claim C. The interpretations about the ability to be assessed are

- meaningful with respect to a particular learning syllabus, an analysis of the abilities needed to perform tasks in the TLU domain, a general theory of language ability, or any combination of these;
- impartial to all groups of test takers;
• **generalizable** to the TLU domain in which the decisions are to be made;
• **relevant** to the decisions to be made; and
• **sufficient** for the decisions to be made.

Claim D. The *assessment records* (scores, descriptions) are **consistent** across different assessment tasks, different aspects of the assessment procedure (e.g., forms, occasions, raters), and different groups of test takers (p. 103, italics and bold in the original; the numbered claims have been changed to letters).

*Figure 1 Construction and Evaluation of an Assessment Use Argument*

![Diagram](image)

The four claims correspond to Messick’s (1989) division between test interpretation (Claims C and D) and test use (Claims A and B). While Claims C and D lend themselves to quantitative data, and Claims A and B to qualitative, it is commonly thought that both methodologies can support the two lines of inquiry, with the data collection method chosen to best suit the inference being made (Bachman, 2009). Indeed, a key strength of argument-based approaches is how they
facilitate the identification of appropriate methods to collect evidence that supports the case being made.

*Figure 2* Example Assessment Use Argument for the CAELD

The purpose of the argument is to meaningfully connect the grounds (e.g., test taker performance, assessment records) to the assessment claim. I have adopted the term grounds over data, the term Bachman and Palmer (2010) use, to avoid confusion with the data collected as part of the research process. Gathering data to form backing allows for the opportunity to identify rebuttals to the warrants, and thereby become evidence to support alternative explanations for the interpretations about test taker performance. I also identified possible rebuttals that require further research; these possible rebuttals were identified as a concern, but there was no information or data to support them. Toulmin (2003) described this process of reasoning within an argument structure as pulling the warrant on either end through backing and rebuttals. Figure
2 depicts the logic of the argument structure. The left-hand side of the figure illustrates how the warrants and the supporting evidence can provide justification for the assessment claim. The figure also demonstrates, along the right-hand side, how evidence collected for the warrants can provide rebuttals for alternative explanations.

**Review of Research Literature**

In answer to Spolsky’s (1992) and Alderson’s (2005, 2007) calls for research into diagnostic assessment, a number of studies have investigated the creation of diagnostic assessments (e.g., Fox, 2009; Jang, 2005; Knoch, 2009; Read, 2008). There is still limited research (with the notable exceptions of Fox, 2009 and Jang, 2009) on how to extend the use of diagnostic assessment feedback into the classroom. This section outlines the research literature as it relates to using a learning-oriented assessment to inform scoring and teaching and thus support students in their acquisition of academic English.

**Scoring.** Research into the development and scoring of diagnostic language assessments has increased ten-fold in the last decade (e.g., Alderson, 2005; Alderson & Huhta, 2005; Huhta, 2010; Jang, 2005; Knoch, 2009). Jang (2005, 2009), in the most comprehensive study to date, investigated the validity of the Next Generation Test of English as a Foreign Language (TOEFL)¹ for an additional diagnostic purpose. In her three-phase mixed methods study, she analyzed the diagnostic potential of the test items and the utility of the feedback in two test preparation courses. Jang included multiple participant groups, 2700 test takers, 7 raters, 3 teachers, and 27 students, and a variety of data collection and analytical techniques including Fusion Modelling, think aloud verbal protocols, interviews, and classroom observations. Fusion modelling sets the parameter estimates for determining mastery and non-mastery. Using an

¹ At the time of Jang’s (2005) study, The Next Generation TOEFL was the name used for the present day TOEFL iBT.

² The participants’ quotations are referenced according to data source (pseudonym, data source,
argumentative-based approach to validation (Kane, 2006), Jang linked the interpretations made from the test to the decisions made by the teachers and students in the classroom setting. The findings of importance for this study were the concurrent nature of the students’ reading processes and the challenges Jang (2005) faced in distinguishing between students with mastery and nonmastery of the test subcomponents. The students used different strategies to arrive at the same correct answer. For example, some of the test items clearly discriminated between masters and nonmasters, but others did not; the items that did not discriminate were either too easy or too difficult. Jang noted the challenges the students encountered when told they had mastery of a subcomponent and noted the implications this information had for their future studying. For instance, if students were identified as a master of a subcomponent, did it mean they should no longer try to improve in that area as an interpretation of the feedback?

While Jang’s (2005) study is comprehensive in its development and implementation of a diagnostic assessment for teachers and students, it only looked at the testing, teaching, and learning of reading. Because reading is typically scored using multiple-choice questions, reading tests are more readily adapted for diagnostic assessment purposes that draw on sophisticated psychometric techniques requiring a large item pool per subcomponent. Reading strategies and diagnosis also have a rich and well-established history in general education, making it easier to identify reading skills and strategies for diagnostic assessments in second-language reading (Alderson, 2005). Further, Jang examined the use of the diagnostic results in a test preparation course for the Next Generation TOEFL. Although Jang did not allude to the students’ motivation specifically, it would be expected that the students’ primary interest in the diagnostic feedback was to improve their test scores on the TOEFL, and not necessarily to improve their English
skills. In addition, Jang’s study did not go into depth into how the teachers or students used the feedback to inform their decisions on how to improve teaching or learning.

Similar research drawing on the quantitative methodology used in Jang’s study has focused on creating diagnostic assessments and been conducted on diagnostic assessments of listening (e.g., Sawaki, Kim, & Gentile, 2009). Yet only a handful of studies have investigated the diagnostic assessment of writing (e.g., Cotos & Pendar, 2008; Knoch, 2009; Llosa, Beck, & Zhao, 2011). Writing is of particular interest for this study since academic writing is a skill that students need to be successful at university (Flowerdew & Peacock, 2001).

The underrepresentation of writing in the research is most likely due to the challenges researchers face in diagnosing writing ability. Alderson (2005) argued that indirect measures should be used to diagnose writing, given the difficulty in accurately identifying students’ subcomponents and the need to provide timely feedback. The example offered by Alderson (2005) is DIALANG, which draws on indirect items that ask students to identify mistakes in text or choose the correct phrase rather than composing an essay. Since DIALANG is offered online without any subjective scoring, feedback is immediate. Alderson’s (2005) book outlines numerous studies that have examined the validity of three skills measured by DIALANG, including writing. While tests such as DIALANG are useful for beginning level students, many have argued that indirect measures are counter to the natural writing process and not representative of students’ future writing tasks (Knoch, 2007a; Weigle, 2002).

Another option for diagnosing writing ability is a task-based approach, in which students compose text in response to a question or, more likely, an essay prompt; while more time-consuming, such tasks mirror those that students would encounter in their future university studies such as writing an argumentative essay as seen in the CAELD. In task-based writing,
there are five factors to consider: the writing task, the performance, the scale, the rater, and the student (McNamara, 1996; Weigle, 2002). While automated scoring is gaining prominence, most essays are still scored by a human rater (Xi, 2010). Research studies investigating direct measures of diagnostic writing ability have considered task type (Llosa, Beck, & Zhao, 2011) and scale (Knoch, 2007a). To date, there appears to be no research investigating raters’ perceptions and knowledge of students’ strengths and weaknesses in writing for the purposes of diagnosing students’ abilities.

Raters are a valuable source of information about students’ writing ability, as they must synthesize the students’ writing ability to a few scores when using analytic scales or, even, to only one score with holistic scales. Analytic scales score a number of individual categories that are then combined for a total score, whereas holistic scales may consider different categories but only provide one, overall score (Hamp-Lyons, 1991). The CAEL and CAELD are scored with a holistic scoring guide, suggesting that raters must make multiple decisions about students’ abilities but only assign one band score (CAEL, 2008). By drawing on rating experience, combined with teaching experience, the rater may provide a critical connection between identifying the students’ skills and providing them with feedback on how to improve. Furthermore, gathering raters’ perceptions of the rating process may help reveal information that would be lost if only using scales; as a result, these perceptions may make the feedback more useful to the learning process.

While there has been limited research investigating the rater perspective on how to diagnose student writing ability, there have been numerous studies investigating rater variability as a way to reduce reliability concerns across raters (intrarater reliability), for the same rater (intrarater reliability), and for rater bias (e.g., Baker, 2010; Barkaoui, 2007). In rating writing,
scoring patterns reflects the leniency or severity that a rater demonstrates when scoring essays (Weigle, 2002). Of interest for this study are the factors that influence raters’ conceptualizations of a text, namely, rating experience (e.g., Barkaoui, 2010; Cumming, 1990), rater reading strategy and approach (Lumley, 2002, 2005; Smith, 2000; Vaughan, 1991), and type of scale and criteria used (Barkaoui, 2007, 2010; Lumley, 2002, 2005; Schaefer, 2008).

Researchers have employed a number of quantitative and qualitative data collection techniques to investigate rater decision-making behaviour, most notably think aloud or verbal protocols. Barkaoui (2010a) investigated the extent to which rating experience influences how a rater uses an analytic or holistic scale. Barkaoui asked 25 raters (11 novice, 14 experienced) to complete 12 think alouds in the convenience of their own home. By coding the think aloud data for decision-making behaviours and interpretation and judgment statements, Barkaoui concluded that the experienced raters were able to develop a deeper conceptualization of the text and identify more features of the essay than novice raters. This ability to have a deeper sense of text qualities is advantageous from a diagnostic assessment perspective, since the experienced rater is able to identify a broader range of students’ strengths and weaknesses.

In a two-phase study, Knoch (2007a, 2007b) drew on previous versions of the AUA described in Bachman (2005) to develop a new scale and then validate it. In the first phase, Knoch developed a new empirically derived analytic scale. To investigate the usefulness of the empirically derived scale, she interviewed 7 raters and administered questionnaires to 10 raters in phase two of her study. Knoch (2007b, 2009) noted that a greater level of detail in the criteria descriptors positively influenced raters’ ability to use a broader range of the scale. The Diagnostic English Language Needs Assessment (DELNA) program chose to adopt the new,
more detailed scale developed by Knoch because of its improved overall reliability and diagnostic potential.

Approaches to rating an essay and rater reading strategies have received considerable interest in recent years. Despite raters’ overall similar approaches to rating, Lumley (2005) and others have noted that the raters were influenced by their initial impressions of the essays and ultimately adopted individual strategies to cope with the borderline essays—those on the cusp of two score levels (Cizek & Bunch, 2006). Smith (2000) found that the six raters in his study adopted one of three reading strategies: (a) “read-through-once-then-scan,” (b) “performance criteria-focused,” and (c) “first-impression-dominates” (p. 171). Drawing on think alouds, Smith observed that raters focused on textual features as well as the scale criteria, particularly when confronted with an essay they found difficult to place on the scale. From a traditional scoring perspective, this inclusion of features outside the scale may cause concerns regarding the reliability and validity of the rating. In fact, it may be important to consider these extra features from a diagnostic assessment perspective as well, as they may illuminate aspects of the writing that are helping or holding the students back in their writing ability.

A predominant challenge in diagnostic assessment is connecting the identified strengths and weaknesses to teaching and learning (Alderson, 2010; Davidson, 2010). A communication bridge is needed to link the diagnostic feedback from an externally developed test to classroom teaching and learning; developing this link will ultimately lead the students to improve their language skills. At the classroom level, this bridge could be established through communication with teachers and students. Outside of a classroom context, the diagnostic feedback can be communicated to the students by other methods. For example, at the University of Auckland, students who are identified as needing support are requested to pick up their DELNA feedback in
person, so the profile can be explained to them (Read, 2008). Before the face-to-face communication of the feedback was implemented, the students were provided the feedback electronically; this practice was revised, as students who most needed support often did not follow through on the feedback and advice to take support courses (Bright & van Randow, 2004). Thus attention needs to be given to how the feedback and advice are delivered to students.

**Teaching.** The Classroom-based Assessment (CBA) literature provides the theoretical grounding for examining how diagnostic assessments are integrated into instruction (e.g., Black & Wiliam, 1998; Leung, 2004; Rea-Dickens, 2001; Sadler, 1989; Wiliam, 2010). My definition of *classroom-based assessment* is broadly in line with Rea-Dickens’ (2008) articulation of the term, which refers to any teacher assessment, whether it is formative or summative in nature. Extending this broad definition, Hill and McNamara (2012) stated that CBA is “any reflection by teachers (and/or learners) on the qualities of a learner’s (or group of learners’) work and the use of that information by teachers (and/or learners) for teaching, learning (feedback), reporting, management or socialization purposes” (p. 396). A key emphasis in Hill and McNamara’s (2012) definition is their attention to both the teachers and students as test users (i.e., assessment agents) with a traditional focus on teaching, learning, and reporting, as well as more non-traditional uses of management and socialization. In addition, the authors broadly defined the possible sources of evidence to be “any actions, interactions or artifacts (planned or unplanned, deliberate or unconscious, explicit or embedded) which have the potential to provide information on the qualities of a learner’s (or group of learners’) performance” (Hill & McNamara, 2012, p. 397; italics in original). The dynamic nature of teaching is reflected in this expansive definition, which includes multiple forms of planned or unplanned assessment opportunities. The definition underscores the idea that an assessment, such as the CAELD, is only one of the many sources of
evidence the teacher is likely to employ as part of her or his pedagogical decision-making process.

Table 3

*Variations of Formative Assessment Characteristics* (McMillan, 2010, p. 43)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Low-level formative</th>
<th>Moderate-level Formative</th>
<th>High-level Formative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evidence of student learning</td>
<td>Mostly objective, standardized</td>
<td>Some standardized and some anecdotal</td>
<td>Varied assessment, including objective, constructed response, and anecdotal</td>
</tr>
<tr>
<td>Structure</td>
<td>Mostly formal, planned, anticipated</td>
<td>Informal, spontaneous, “at the moment”</td>
<td>Both formal and informal</td>
</tr>
<tr>
<td>Participants involved</td>
<td>Teachers</td>
<td>Students</td>
<td>Teachers and students</td>
</tr>
<tr>
<td>Feedback</td>
<td>Mostly delayed (e.g., give a quiz and give students feedback the next day) and general</td>
<td>Some delayed and some immediate and specific</td>
<td>Immediate and specific for low-achieving students, delayed for high-achieving students</td>
</tr>
<tr>
<td>When done</td>
<td>Mostly after instruction and assessment (e.g., after a unit)</td>
<td>Some after and some during instruction</td>
<td>Mostly during instruction</td>
</tr>
<tr>
<td>Instructional adjustments</td>
<td>Mostly prescriptive, planned (e.g., paced according to an instructional plan)</td>
<td>Some prescriptive, some flexible, unplanned</td>
<td>Mostly flexible, unplanned</td>
</tr>
<tr>
<td>Choice of task</td>
<td>Mostly teacher determined</td>
<td>Some student determined</td>
<td>Teacher and student determined</td>
</tr>
<tr>
<td>Teacher-student interaction</td>
<td>Most interactions based primarily on formal roles</td>
<td>Some interactions based on formal roles</td>
<td>Extensive, informal, trusting, and honest interactions</td>
</tr>
<tr>
<td>Role of student self-evaluation</td>
<td>None</td>
<td>Tangential</td>
<td>Integral</td>
</tr>
<tr>
<td>Motivation</td>
<td>Extrinsic (e.g., passing a competency test)</td>
<td>Both intrinsic and extrinsic</td>
<td>Mostly intrinsic</td>
</tr>
<tr>
<td>Attributions for success</td>
<td>External factors (teacher; luck)</td>
<td>Internal, stable factors (e.g., ability)</td>
<td>Internal, unstable factors (e.g., moderate student effort)</td>
</tr>
</tbody>
</table>
CBA is broadly defined to include summative and formative assessments. However, one of the goals of the study is to examine the extent to which the CAELD informed teaching and learning—in other words, its potential as a formative assessment tool. While I defer to Wiliam’s (2011) notion of formative assessment as an evidence-based practice informing the teaching and learning process, it is also important to highlight the crossover in terminology in the educational assessment literature. The distinctions between summative and formative assessment can also be defined as assessment of, for, and as learning (see Rea-Dickens, 2008). Assessment of learning reflects the summative decision-making about what students have learned. While assessment for and as learning fit within formative assessment, the former focuses on determining next steps for teaching and learning, whereas the latter is concerned with developing a student’s metacognitive awareness about his or her learning through self-assessment (Black & Wiliam, 1998; Earl, 2003).

Thus there is not one type of formative assessment but rather a range of assessments encapsulated within the term. McMillan (2010) described this range as low-, moderate-, and high-level formative assessments. Low-level formative assessments only provide feedback to teachers who may or may not consider the feedback. At the other end of the spectrum, high-level formative assessments are fully integrated into instruction. Teachers and students are active participants and the classroom provides a supportive and encouraging environment in which students are able to learn about themselves as learners. This definition differs from Wiliam’s (2010), which stated that the feedback only needs to be considered to be formative. McMillan distinguished the three levels by 11 characteristics seen in Table 3. A characteristic missing from McMillan’s characterization of formative assessments is the content of the assessment. It is implied that there will be a strong connection between the content of the assessment and the curriculum. If the content is not aligned, then the feedback is inappropriate and meaningless to
both the teacher and students. Leung (2005) and Wiliam (2001) and others have investigated the alignment between CBA test content and the curriculum. Specific to diagnostic assessments, Jamieson, Grgurovic, and Becker (2008) developed diagnostic assessments based on course textbooks to ensure the test content mapped onto the curriculum taught. Formative assessments are defined by their use, feedback cycle, and interaction between teachers and students.

A number of factors can influence how formative assessment is employed by teachers and received by students. Contextual factors include, but are not limited to, course content, policy, sociocultural differences between students and teachers, and classroom environment (Hill & McNamara, 2012; McMillan, 2010). Teachers occupy a major role and may differ in ability and perceptions of formative assessment (Fox, 2009; Jang, 2005, 2009). Students are likely to differ in motivation, self-regulation, ability, and age (Jang, 2005, 2009).

Shepard (2000) averred that fostering a learning culture is critical for engaging students in assessment practices that encourage reflection on how to improve. In such a setting, assessment supports instruction rather than constraining it by requiring that the assessment must direct the instruction. Despite these ideals, educational settings often place restrictions on teachers regarding the content and type of assessments they can use and the degree to which they can emphasize assessment within instruction and learning (Davison, 2007; Davison & Leung, 2009; Rea-Dickens, 2001, 2007).

Diagnostic assessments are typically developed external to the classroom environment and/or include high-stakes outcomes (Huhta, 2007; Popham, 2009). Using standardized assessments—ones that align to curriculum standards and to models of language proficiency—strengthens the validity of the interpretations that teachers form about their students’ abilities (Jamieson, Grgurovic, & Becker, 2008). Thus teachers can more accurately and confidently
identify students’ language needs for diagnostic, formative, or summative purposes. Integrating this information should support the positive use of assessment within classroom teaching and learning and, ultimately, contribute to a learning culture that values data-based learning (Black et al., 2003; Shepard, 2000). To enhance the match between the results and instruction, it is likely that teachers will need to engage in the diagnostic assessment process by personalizing the profile or scoring a portion of the assessment. The time required to adjust course materials and work with students on a one-to-one basis poses an additional constraint for teachers (Black et al., 2003). Such constraints may limit teachers’ pedagogic flexibility when adopting the diagnostic feedback, as well as their ability to take ownership for the overall process and create a supportive assessment culture.

Xu and Liu (2009) captured one teacher’s assessment knowledge and practice using narrative inquiry and noted how the teacher personalized her practice. The authors formed the three narratives based on two semistructured interviews, personal discussions and communications, and personal records (student evaluations, assessment and teaching plans, and any notes the teacher wrote about her assessment practice). Through an in-depth analysis and description, Xu and Liu demonstrated the individualized and personal approach one teacher took with her assessment practice.

While the contextual factors influence how CBA takes shape, it is really the teachers who embody the pivotal role as agents mediating the process (Rea-Dickens, 2004). Teachers decide which aspects of the assessment results to integrate into their teaching, and when and if to present the feedback to students (Rea-Dickens, 2008). When designing a course with the overt goal of including formative assessments, teachers must consider the purpose, criteria, process, and student involvement of the assessment (Davison & Leung, 2009; Rea-Dickens, 2001).
Even though teachers play a pivotal role in integrating assessments and feedback into instruction, limited research has examined the process of CBA compared to the outcomes (Hill & McNamara, 2012). In response to the gap in literature, Hill and McNamara examined the evidence used and interpretations made by teachers in Indonesian-as-a-foreign-language classrooms. In their study, the authors examined two foreign language classrooms in Australia with students aged 11–13. Recognizing the challenges in gathering teachers’ interpretations and uses of classroom assessment data, Hill and McNamara triangulated interviews and classroom observations. The authors identified the four main processes of planning, framing assessments for students, conducting the assessment, and using the results in their study. Similar to Bell and Cowie’s (2001) study, the teachers in Hill and McNamara’s study utilized both spontaneous and structured planning.

Hill and McNamara (2012) noted the challenges the teachers had in framing the assessments for students, given that the students in their study were young adolescents and that the assessments were often embedded in instruction. Nonetheless the authors, along with Rea-Dickens (2006), advocate strongly for students to be actively engaged, noting such engagement to be an attribute of effective CBAs. To identify the type of engagement by students, Torrance and Pryor (1998) contrasted divergent and convergent questioning from teachers. Divergent questioning focuses on discrete knowledge and directs students to the desired responses, while convergent questioning is open and exploratory, leading to deeper learner engagement (Torrance & Pryor, 1998). Similar to those in Torrance and Pryor’s (1998) study, Hill and McNamara’s (2012) teachers used assessment for a variety of purposes, including socialization and management, as a means to introduce the students into the local conventions of assessment as well as to move a lesson along as part of managing a lesson. The processes of CBA are
inherently complex and layered, encompassing the pedagogic act of determining students’ needs, teaching to those needs, and even classroom management.

Teachers’ perceptions of the test or assessment data proved critical in Fox’s (2009) study in an EAP program. Fox collected interview data from four teachers as well as email and informal chat data from the program coordinators. Fox observed the teachers’ uptake of the CAELD feedback as passive or negative and examined the impact of a policy change allowing students to be placed into the university’s EAP program based on scores from various tests such as TOEFL or IELTS. Given the possible continuous and negative impact the teachers might have experienced during Fox’s study, it is difficult to discern the extent to which the policy changes impacted the teachers’ perceptions and use of the CAELD feedback. Further, Fox did not conduct any classroom observations to link the actual use of the feedback to the teachers’ statements in the interviews.

Research has demonstrated that teachers’ previous experience and ability to effectively interpret assessment information is a critical element in successfully using formative assessment (Davison, 2004; Harlen & Winter, 2004; Rea-Dickens, 2006). This ability factor is, in fact, so critical that Wiliam (2010) and others demanded that educational training programs include courses instructing new teachers on how to use formative assessment. The ability factor that is most appropriate for this study is teachers’ diagnostic competence. Edelenbos and Kabanek-German (2004) argued that teachers require a high level of diagnostic competence to facilitate fair and accurate assessment of students. Edelenbos and Kabanek reanalyzed videos of classroom observation data to formulate their model of diagnostic competence. In their paper, Edelenbos and Kabanek-German (2004) defined diagnostic competence as “the ability to interpret students’ second or foreign language growth, to skilfully deal with assessment materials
and to provide students with appropriate help in response to this diagnosis” (p. 260). The authors elaborated that they viewed diagnostic competence as an attribute of teachers and not of diagnostic assessments. In addition to the ability that teachers have to interpret assessment information, teachers’ perceptions and values attributed to assessment and teaching influence how they integrate the diagnostic or formative assessment information into their lessons (Colby-Kelly & Turner, 2007; Fox, 2009; Jang, 2005, 2009; Leung, 2004).

Broader than the skill of diagnostic competency is assessment literacy—the knowledge base that individuals involved with language testing have about issues surrounding assessment, such as validity, reliability, and test fairness (Taylor, 2009). According to Taylor, assessment literacy has gained prominence in recent years due to the increased use of large-scale assessments for accountability purposes and the recognized need to enhance understanding of language-assessment development and use across a number of stakeholder groups, one of which is teachers engaged in CBA. In the classroom context, assessment literacy contributes to teachers’ competence in using assessments for a variety of uses, such as the formative uses described above—including teaching, management, and socialization—as well as more summative uses.

Davison (2005) noted that teachers often struggle with distinguishing between the different types of assessment, for example, the purposes and uses of summative and formative assessment. This limited understanding is not surprising, since most applied linguistics programs often offer one course on language testing (Taylor, 2009). Further, McNamara and Roever (2006) underscored the fact that language-testing courses are often narrow in scope, so that the students, as future teachers, are not exposed to the range of complex, larger issues of social consequences of test use. In line with Taylor (2009) and McNamara and Roever (2006), Fulcher
(2012) proposed that language-assessment literacy should be broadly conceptualized based on the contexts, principles, and practices of using assessments. Thus diagnostic competence and assessment literacy are likely to be essential components of integrating diagnostic assessment into instruction for beneficial consequences.

In summary, the CBA literature provides critical insights into the ways in which diagnostic assessment feedback might be integrated into classroom instruction, the factors contributing to a teacher’s interpretation of the feedback, the notion of fostering a positive learning culture, and the need for alignment between test content and curriculum. While the teacher is an important stakeholder in the formative use of a large-scale assessment for diagnostic purposes, students are perhaps the most important stakeholders in the equation; the goal of the process is, after all, to help improve their learning.

Learning. By the time students enter university, they are well-trained students with at least 11 years’ experience in an educational system. Throughout their educational careers students are often passive in the decisions surrounding assessment of their ability, gaining extensive knowledge and experience of having assessments done to them rather than with them (Boud, 2007). This unbalanced experience with summative assessments would be more pronounced for students from education systems with accountability frameworks that overshadow formative assessment (Berry, 2011). In fact, some students might never have had any experience with formative assessment. It is unreasonable to expect students who have had no previous experience with formative assessment to become active and knowledgeable participants. There is a growing body of literature examining students’ use of assessment information for their own learning in general education contexts (e.g., Andrade, 2010; Nicol & Macfarlane-Dick, 2006; Taras, 2010).
As experienced test takers, students have first-hand knowledge and insight into the testing process. However, this personal account of a test, sometimes thought of as face validity, has been typically discounted because the layperson did not have the knowledge or skills to adequately examine the test (Davies & Elder, 2005). The student perspective as a valuable data source first started gaining traction in the language testing field with Alderson and Wall’s (1993) and Cheng’s (2005) washback studies. Since then, researchers have drawn on the student perspective to investigate test impact (e.g., Cheng, Andrews, & Yu, 2011), test construct (e.g., Xi, 2010), and identity (e.g., McNamara & Roever, 2006). Despite the increase in the number of studies focused on the student perspective, there have been few studies that have investigated students’ interpretations of test feedback within the language assessment field.

Within an assessment for learning context, Colby-Kelly and Turner (2007) sought to explain the nature of formative assessment through the teacher and student perspectives. In their study, Colby-Kelly and Turner adopted an earlier version of the AUA (Bachman, 2005) to examine the students’ perceptions about the types of formative feedback they received on speaking tasks in an EAP program. Drawing on questionnaires and interviews with the students, the authors noted that the students’ perceptions of the teacher’s formative feedback varied from motivating to unmerited and untrustworthy. The authors proposed the concept of the assessment bridge since assessment was the link between teaching and learning.

Another important factor influencing students’ perceptions of diagnostic assessment is in comparison to their self-assessment. To investigate students’ perceptions about the Vocabulary Size Placement Test (VSPT) of DIALANG, Huhta (2007) administered 533 questionnaires, with Likert and open-ended questions, to university students in Finland and Germany. Huhta observed a range of reactions to the feedback, most notably with the qualitative responses. The lowest-
scoring students disagreed with the feedback based on their perceived ability. In addition to looking at student reactions to feedback, the author also examined the students’ understanding of the construct. Huhta attributed the students’ familiarity with the test to their understanding of the construct. More recently, Yin, Sims, and Cothran (2012) interviewed five students and administered 68 questionnaires to students at a Chinese university to examine the usefulness of diagnostic test items’ feedback. Similar to Huhta’s findings, the higher proficiency students perceived the feedback as being more useful.

Further, student perceptions of a test may be influenced by the test topic (Fox, 2004), the testing context (Brown, 2008; He & Shi, 2008), and the length of time given to take the test (Knoch & Elder, 2010). Beyond the testing event, students’ reactions to feedback may influence the degree to which they draw on the information to direct their studying habits. According to Alderson (2005), the specificity and timing of feedback should impact whether or not students attend to the feedback given. Overall, research has shown that examining perceptions is useful for understanding the intricacies of student experiences and perceptions about a test.

Diagnostic assessment is increasingly described as a potentially beneficial tool for classroom instruction—serving as a mediating tool between student learning and teaching (Alderson, 2005). The diagnostic feedback obtained can provide information about what the students are able to do and what they cannot do, thus allowing teachers the opportunity to tailor their instruction to students’ specific needs.

The Assessment Use Argument for the CAELD

I adopted the AUA to guide this validation study of the CAELD for use in an EAP classroom. I now describe Assessment Claim A, Beneficial Consequences, the desired outcome resulting from the use of the CAELD (please see Figure 1 for a depiction of the construction and
Appendix F outlines the four assessment claims and 27 warrants related to beneficial consequences, decisions, interpretations of students’ assessment performance and feedback, and consistency of the assessment records. I modelled the 27 warrants on the example warrants provided in Bachman and Palmer (2010). The individual warrants have a letter and number format (e.g., A.1). The letter refers to the assessment claim and the number to the warrant within that claim. Based on the above review of the literature, three stakeholder groups were identified as being central to the diagnostic process in this study: raters, teachers, and students. To make the claims and warrants as specific as possible, I sometimes refer to the participants in the study: the two raters, one teacher, and the teacher’s 47 students.

**Beneficial consequences.** The origin of the beneficial consequences assessment Claim A connects back to Kunnan’s (2010) notion of beneficence in his test fairness framework; for him, a test should “bring about good to a society… [as well as] not inflict harm to society” (p. 184). The characterization of good or harmful to a society is determined in the local context; as such, the beneficial consequences assessment claim in this study is centred on the teacher and students. The five beneficial consequences warrants specify that the CAELD feedback should support both the teacher in her or his instruction in an intermediate-level EAP course and the students’ learning within the course, while not hindering instruction or learning.

**Decisions.** Claim B makes explicit that the decisions derived from the interpretations of the CAELD (Claim C) reflect the values of the teacher and students and are equitable for all of the students. The first two warrants qualify the value sensitivity dimension for the teacher and students of the decisions Claim (B.1-B.2) and whether or not the decisions made are equitable for all of the students (B.3-B.4).
**Interpretations.** Assessment Claim C seeks to determine that the information gathered about the students is based on a well-defined construct, fair to all student groups, suitable for the decisions needed, and sufficient enough to have confidence in the inferences (Bachman & Palmer, 2010). This assessment claim is the most complex with 5 warrant categories (i.e., meaningfulness, impartiality, generalizability, relevance, and sufficiency) and 17 warrants in total.

**Meaningfulness.** The six meaningfulness warrants elaborate on the content validity and construct validation of the interpretations assessment Claim (C.1 – C.4) and expand on the transparency of the construct and assessment report for stakeholders (C.5 – C.6). Therefore, the first half of this warrant category refers to the appropriateness of the defined construct for the intended use; the second half highlights the obligations of the test developers—in the case of this study, the researcher (myself) and the teacher—to communicate the construct and results to the stakeholders, the teacher and students. The Target Language Use (TLU) domain for the CAEL are first-year university courses, while the TLU domain for the CAELD is the EAP program. The two domains are similar enough that the CAEL is used for purposes of placement into the program (see Chapter 1). The existing CAEL test specifications and research related to academic English provided the basis for the diagnostic subcomponents (Fox, 2009).

**Impartiality.** The five impartiality warrants specify whether or not the content or format is free from bias for specific subpopulations (C.7), is culturally sensitive to all possible students taking the test (C.8), is accessible to all test takers in information and physical ability (C.9-C.10), and results in meaningful interpretations of all test takers regardless of background (C. 11). Bias within this impartiality warrant category relates to systematic group differences in test performance that do not reflect differences in ability (Kunnan, 2004). Since the CAELD is the
same as the CAEL in content and format, the latter having been vetted for bias across a range of test populations, it is unlikely that warrants C.7-C.8 are of any concern. Warrant C.10 reflects the need for all students to be able to access how the feedback was created—in this case, through the learner profiles.

**Generalizability.** The two generalizability warrants qualify the connection between the test tasks and the TLU tasks (C.12), and the degree that the scoring and recording procedures are similar to the evaluation in the TLU domain (C.13). This warrant category is connected to the meaningfulness warrants, in that the meaningfulness warrants make explicit the construct definition, while the generalization warrants examine the degree to which the test tasks do, in fact, correspond to the defined TLU tasks.

**Relevance.** The three relevance warrants specify that the interpretations about students’ writing ability are able to map on directly to specific next steps for student learning (warrant C.14), and relevance of the feedback to the teacher’s course instruction and students’ learning (C.15-C.16). Once again there is a connection to the meaningfulness category. The raters were identified as the data source for warrant C.14, since they could connect the observations about the students’ test performance to direct feedback.

**Sufficiency.** The single sufficiency warrant specifies the information provided about students’ test performance is ample enough for the decisions being made based on the test developers’, test users’, and stakeholders’ comfort levels (C.17). Bachman and Palmer (2010) acknowledge that the stakes attributed to the decisions will help determine the sufficiency of the information needed.

**Assessment Records.** The final assessment claim qualifies the reliability of the scoring process. There are two main concerns with the reliability of assessment records: random and
systematic error (Haladyna & Downing, 2004). Warrant D.1 is concerned with the need to identify possible sources of systematic error.

**Summary**

In this chapter on validity theory and frameworks, I described how validity theory, and in turn validation, is centred on test use and the underlying interpretations substantiating that use. A valued component of the validation process is examining the social consequences stemming from test use and the social value attributed to the results (Kane, 2009; Messick, 1989; McNamara & Roever, 2006). Within such complex dynamic education systems, multiple methods from a range of stakeholder groups are seen as necessary to arrive at a rich understanding of the test use (Cheng, 2008; Fulcher, 2009; Shohamy, 2001). Further, argument-based approaches have proven meaningful for guiding validation research that draws on a range of methodologies (Bachman, 2009).

Therefore, drawing on contemporary validity theory, this study aims to evaluate the extent to which the CAELD is appropriate for diagnostic use in an EAP program. The Assessment Use Argument (Bachman & Palmer, 2010) was adopted in this study to guide the validation of the CAELD, as specified by four assessment claims that were further qualified by 27 warrants.

In this chapter, I also reviewed the research literature related to raters’ scoring of academic writing for diagnostic and non-diagnostic purposes, teachers’ use of assessment results to inform instruction, and finally students’ perceptions and use of feedback. I concluded this chapter by outlining the AUA constructed to examine the appropriateness of the CAELD for diagnostic purposes in an EAP program.
Chapter 3: Methodology

The purpose of this study is to examine the appropriateness of the Canadian Academic English Language Diagnostic (CAELD) assessment for diagnostic purposes in an English for Academic Purposes (EAP) program from the perspectives of three stakeholder groups who were engaged in the processes of scoring, teaching, and student learning. I employed a case study methodology to arrive at an integrative understanding of the interpretations, decisions, and uses of the CAELD feedback from these three vantage points. As such, I decided to first examine each stakeholder perspective independently, using three concurrent phases on the raters, teacher, and students to provide an in-depth examination of the phenomenon of interest—the use of the CAELD in an EAP program in one Canadian university. Subsequently, the findings of the three phases were considered as a whole to offer a comprehensive evaluation of the use of the CAELD for diagnostic purposes to inform classroom instruction and support student learning.

Qualitative Case Approach

A case study method was employed to gain an in-depth understanding of the phenomenon of the use of the CAELD in a specific instructional setting. Yin (2009) outlines two qualities that separate case studies from other forms of social science research, particularly those using qualitative methods. According to him, a case study

(a) investigates a contemporary phenomenon in-depth and within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident, [and]

(b) copes with the technically distinctive situation in which there will be many more variables of interest than data points, and as one result relies on multiple sources of evidence, with data needing to converge in a triangulating fashion; and as another
result, benefits from prior development of theoretical propositions to guide data
collection and analysis. (p. 18)

This study encompasses these two qualities. First, the boundaries surrounding the use of
the CAELD and the EAP program context are indistinguishable spatially and temporally. The
administrative office for the CAEL is housed within the same building as the EAP program. In
scoring the CAEL essays, raters are advised to ground their decisions in the scoring guide as well
as consider the student and her or his placement into the EAP program (CAEL, 2008). This close
physical and conceptual proximity likely influences raters’ scoring of the CAEL essays.
Boundaries are similarly blurred for teachers in the EAP program, as they are familiar with the
CAEL. Finally, the student, who may have previous experience with the CAEL as a language
proficiency test, must make decisions on how the results will influence her or his learning. Thus
the three groups of raters, teachers, and students are key stakeholders and decision makers in
how the CAELD is used within an EAP program.

Second, to arrive at a comprehensive understanding of the uses of a retrofitted language
proficiency test, multiple sources of evidence were required. By including three key stakeholder
perspectives, a case study approach revealed interactions and processes of the situation that
might have otherwise been invisible to large-scale studies or a singular vantage point (Yin,
2009). Thus this was an explanatory case study (Yin) with the underlying purpose being to
understand how a large-scale language proficiency test, such as the CAEL, can be used for
diagnostic purposes (CAELD) in a learning setting—specifically an EAP program. The
theoretical framework outlined in Chapter 2, Assessment Use Argument, guided the design of
the data collection strategies and analysis for investigating this additional purpose.
The rationale for a single case was due to its unique context. Through their intimate knowledge of the EAP program, the raters were able to offer both a rare perspective of students’ learning needs and advice on how the students could improve. Another unique quality of the case was the fact that the teacher was able to draw on instructional support atypical of most EAP teaching classrooms. Further, the teacher was able to modify the format and content of the learning profiles created. The instructional support and autonomy afforded to the teacher provided a special opportunity to examine the use of a large-scale assessment for diagnostic purposes when adapted to an individual teaching style and classroom. It was expected that some of the students would have previous experience with taking the CAEL in order to be placed into the EAP program, and thus might be more amenable to using the feedback.

This case was bounded by place, time, and perspective. The study examined the use of the CAELD in one term of study at one EAP program (December 2009-April 2010). The overall design was a single case with three embedded units of analysis: (a) two raters, (b) one teacher, and (c) the 47 students taught by the teacher. The main unit of analysis was the use of the CAELD in the EAP program. The three key stakeholders were chosen to gain a multi-faceted, but centered, perspective of validating a large-scale assessment retrofitted for diagnostic purposes. Other meaningful viewpoints not included were the test developers, program coordinators, and the students’ future professors of the content courses. However, it was important to focus on and limit the perspectives to those with the closest proximity to the three main processes of diagnostic assessment: scoring, teaching, and learning. In line with case study methodology, this research included multiple data sources from each stakeholder group in order to gather a rich and thick description about the context and multiple interpretations of validating the CAELD.
Participants

Three participant groups were involved in this study: two raters, one teacher, and 47 students—all from within an EAP program at one Canadian university. This research received clearance from Queen’s University’s General Research Ethics Board and from the ethics board at the university where this study took place (see Appendix G). Both boards comply with the Tri-Council policies on research with human subjects. Prior to data collection, all participants were given their respective letters of information and signed the corresponding consent form. Appendix G also includes my certificate of completion for the online Course in Human Research Participant Protection (CHRPP). Each of the participant groups is described in more detail below.

Raters. Two experienced female CAEL raters participated in Phase 1. The raters were assigned pseudonyms, Catherine and Isobel. Catherine started as a teacher in the EAP program 17 years ago and began rating CAEL essays a year or so later, but took a four-year break from working for the CAEL. At the time of the study, Catherine had returned to being a full-time CAEL rater for three years. Isobel had 10 years of full-time rating experience with the CAEL essay. Over those 10 years, she also taught CAEL preparation and summer EAP courses.

Teacher. The sole participant for Phase 2 was a female EAP teacher, to whom I refer as Diane. Teaching was Diane’s second career. After working in banking and finance for more than 10 years, Diane returned to school to complete her Master’s of Applied Language Studies. She started teaching in the EAP program in 2004, upon receiving her degree. Therefore, at the time of this study, she had six years of teaching experience in the EAP program.

Diane had previous experiences developing diagnostic assessment tasks and receiving diagnostic feedback about her students’ language ability. In fact, this study offered Diane her
fourth opportunity to use diagnostic feedback to inform her teaching. She had previously used results from a test for international teaching assistants at the university, which was piloted with her class, so she already had experience using language-proficiency test results diagnostically: “I was the marker and I used the results to explicitly guide my teaching. It was a smaller group, so that was a unique situation” (Diane Interview 12). In addition, Diane had two experiences of receiving CAEL diagnostic feedback about her students’ language ability. In both cases with the CAELD, Diane indicated that she felt that she had never really implemented the diagnostic feedback into her instruction, nor had she given the feedback to the students for their use.

Beyond her experience teaching EAP and receiving diagnostic feedback, Diane had previously worked as a rater, scoring CAEL essays and marking question booklets. Her experience working for the CAEL gave her a better understanding of the test, and the diagnostic subcomponents that would be most appropriate for informing her teaching and, from her perspective, for the students’ learning.

Diane was responsible for teaching the two intermediate-level EAP course sections (Section A and Section B3). Both sections had two morning classes every week, on Mondays and Wednesdays (Section A), and Tuesdays and Thursdays (Section B). All classes were scheduled from 8:30 to 11:30, with a 15-minute break. The CAELD was given during the second class for both sections. Both sections observed in this study had the same course number and level, with identical course outlines and timelines, and nearly identical lesson plans. When Diane was unable to cover all of the material in one class, it would carry over to the next class, though the other section was perhaps able to cover all of the material in a single class. Diane did notice an

2 The participants’ quotations are referenced according to data source (pseudoynm, data source, and number).
3 The labels of Section A and Section B do not correspond to the actual names used for the course sections.
initial difference between the two sections. In section A, she felt the students were more engaged. “They all come, they are on top of things, they are ready to go, and they participate in the class. It's great.” With section B, she felt that more of the students were likely to be late, “When I get there, there is like maybe 10 people there, 15 minutes later a few more straggle in” (Diane Interview 1). Overall she did seem to express enthusiasm for teaching both sections.

**Students.** There were 47 students who agreed to participate in the study. Not all of these 47 students completed the three reflections; however, all of them did complete at least one for a total of 123 Reflections. In all, 38 completed Reflection One; 42 completed Reflection Two; and 42 completed Reflection Three. All the students (29 male and 13 female) completed the demographic information on Reflection Three. The students’ ages ranged from 17 to 52, with a median and mode of 20. Six first languages were reported: 20 students spoke Chinese (Mandarin and Cantonese); 16 Arabic; 2 Kurdish; 3 Farsi; 1 Vietnamese; and 1 Bengali. The students reported nine home countries: China (20 students), Saudi Arabia (12), Iran (3), Libya (2), Egypt (1), Bangladesh (1), Vietnam (1), Palestine (1), and Afghanistan (1). All of the students, in addition to taking the intermediate-level EAP course, were enrolled concurrently in at least one content course at the university. All of the participants were assigned a number to ensure, to the extent possible, that their identity remained confidential. Of the 47 students, 9 participated in two interviews each.

Table 4 provides an overview of demographic information of these students. To maintain continuity between the quotes in the Phase 3 results chapter, I assigned pseudonyms, rather than numbers, for the students interviewed. There is an uneven balance between the number of males interviewed versus females, but it does reflect the gender distribution for the two courses. Females account for 30% of the 42 students who completed Reflection 3, and 22% of the 9.
### Table 4

**Demographic Information of Students Interviewed**

<table>
<thead>
<tr>
<th>Participant Pseudonym</th>
<th>Gender</th>
<th>Age</th>
<th>First Language</th>
<th>Home Country</th>
<th>Length of time in Canada (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Douglas</td>
<td>M</td>
<td>22</td>
<td>Chinese</td>
<td>China</td>
<td>4</td>
</tr>
<tr>
<td>Dennis</td>
<td>M</td>
<td>31</td>
<td>Arabic</td>
<td>Libya</td>
<td>24</td>
</tr>
<tr>
<td>Stephen</td>
<td>M</td>
<td>21</td>
<td>Arabic</td>
<td>Saudi Arabia</td>
<td>25</td>
</tr>
<tr>
<td>Laura</td>
<td>F</td>
<td>52</td>
<td>Farsi</td>
<td>Iran</td>
<td>40</td>
</tr>
<tr>
<td>David</td>
<td>M</td>
<td>19</td>
<td>Arabic</td>
<td>Saudi Arabia</td>
<td>9</td>
</tr>
<tr>
<td>Don</td>
<td>M</td>
<td>17</td>
<td>Arabic</td>
<td>Egypt</td>
<td>4</td>
</tr>
<tr>
<td>Neil</td>
<td>M</td>
<td>23</td>
<td>Chinese</td>
<td>China</td>
<td>3</td>
</tr>
<tr>
<td>Elaine</td>
<td>F</td>
<td>20</td>
<td>Bengali</td>
<td>Bangladesh</td>
<td>30</td>
</tr>
<tr>
<td>Kevin</td>
<td>M</td>
<td>21</td>
<td>Arabic</td>
<td>Saudi Arabia</td>
<td>14</td>
</tr>
</tbody>
</table>

### Instruments

The instruments identified as being the most appropriate methods were based on the warrants developed out of the AUA for the CAELD. These warrants in turn informed facilitating questions that helped answer the three main research questions. The data collected using the instruments formed the backing to support the warrants for the assessment claims. For an overview of the 3 research questions and 12 facilitating questions across the 3 phases along with the corresponding methods please see Table 5. The instruments are discussed in detail in the phases sections below.

**Phase 1.** Two instruments were used in Phase 1: (a) think aloud protocol and (b) interview protocol.

**Think Aloud Protocol.** The think aloud protocol was used to elicit raters’ perceptions of students’ strengths and weaknesses in CAELD writing. According to Green (1998), a think aloud protocol represents the data collected about a participant when he or she is asked to think aloud.
The protocol refers to the “utterances made as an individual carries out a single task” (Green, 1998, p. 1). These utterances are taken to be a truthful account of what an individual attends to (or had attended to) when completing a task. The use of think aloud protocols in language-learning and language-testing research is increasingly recognized as a valid and effective method to capture the assessment process (Green; Lazaraton & Taylor, 2007). More specifically, think aloud protocols have been used extensively to examine the validity of the scoring process of language tests by asking raters to think aloud while they are marking test items, tasks, or essays (Alderson & Banerjee, 2002; Ducasse & Brown, 2009; Lumley, 2002). Green lists considerations that a researcher should attend to before and after the think aloud sessions. Green states that preparing clear instructions, briefing with the participants beforehand, and practising the technique and task are essential to setting up a meaningful think aloud session.

The think aloud protocol used in this study had two key aims: (a) eliciting raters’ identification of students’ strengths and weaknesses of their CAELD essays, and (b) connecting the identified strengths and weaknesses to meaningful feedback on how to help students to improve. Specifically, the raters were asked about the features of the writing (e.g., development of ideas, accuracy of structure, organization), based on the rating scale that placed the essay in a particular score level. See Appendix H for the think aloud protocol given to the raters.

Rater Interview Protocol. Interviewing offered a meaningful approach for capturing the raters’ perspectives of scoring a CAELD essay. Moreover, interviews were useful for identifying information that could not be obtained through direct observation (Patton, 2002). To help guide and structure the interviews, protocols were developed. The interviews were semistructured in nature. When needed, the participants were asked additional questions to clarify their responses.
Table 5

Research and Facilitating Questions and Corresponding Data Sources

<table>
<thead>
<tr>
<th>Research and Facilitating Questions</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Phase 1 (Chapter 4)</strong></td>
<td></td>
</tr>
<tr>
<td>1. What are the features of CAELD scoring that enable it to be used for diagnostic purposes?</td>
<td></td>
</tr>
<tr>
<td>1.1. How did the raters pinpoint the diagnostic subcomponents in the students’ writing?</td>
<td>Think aloud protocols, Interviews</td>
</tr>
<tr>
<td>1.2. What were the dimensions of the diagnostic subcomponents, as observed by the raters?</td>
<td>Think aloud protocols, Interviews</td>
</tr>
<tr>
<td>1.3. What was the range of subcomponents observed for selected essays, as observed by the raters?</td>
<td>Think aloud protocols</td>
</tr>
<tr>
<td>1.4. What advice did the raters offer to the students on how to improve their academic writing skills?</td>
<td>Think aloud protocols, Interviews</td>
</tr>
<tr>
<td><strong>Phase 2 (Chapter 5)</strong></td>
<td></td>
</tr>
<tr>
<td>2. What are the teaching conditions that enable the CAELD results to be used for diagnostic purposes?</td>
<td></td>
</tr>
<tr>
<td>2.1. How did the teacher incorporate the diagnostic feedback into her course instruction overall and for individual students?</td>
<td>Individual interviews, Document analysis, Classroom observations</td>
</tr>
<tr>
<td>2.2. How did the teacher interpret the assessment information as diagnostic feedback?</td>
<td>Interview data, Classroom observations, Document analysis</td>
</tr>
<tr>
<td>2.3. What aspects of the diagnostic feedback mapped onto the teacher’s overall course planning?</td>
<td>Interview data, Document analysis</td>
</tr>
<tr>
<td>2.4. How did the decisions made and the diagnostic feedback map onto the teacher’s values about teaching, learning, and assessment?</td>
<td>Individual interviews, Classroom observations, Document analysis</td>
</tr>
<tr>
<td><strong>Phase 3 (Chapter 6)</strong></td>
<td></td>
</tr>
<tr>
<td>3. What are the learning conditions and students’ perceptions that enable the students to use the CAELD and its results to support their learning?</td>
<td></td>
</tr>
<tr>
<td>3.1. How did the students use the diagnostic feedback to benefit their learning?</td>
<td>Interviews, Reflections</td>
</tr>
<tr>
<td>3.2. How did the students interpret the CAELD feedback?</td>
<td>Interviews, Reflections</td>
</tr>
<tr>
<td>3.3. How did the diagnostic feedback map onto the students’ learning goals?</td>
<td>Interviews, Reflections</td>
</tr>
<tr>
<td>3.4. How did the diagnostic feedback map onto the students’ values about learning?</td>
<td>Interviews</td>
</tr>
<tr>
<td>3.5. What were the students’ test taking experiences of the CAEL?</td>
<td>Interviews</td>
</tr>
</tbody>
</table>
There were four main points of interest for the rater interview protocol:

1. *Raters background information* [RBI]⁴: experience as a CAEL rater and ESL/EAP teacher;

2. *Rater perceptions* [RP]: raters’ perceptions of the CAELD writing, and the advantages and limitations of using it for diagnostic purposes;

3. *Perceived strengths and weaknesses* [PSW]: insight into the raters’ perceived strengths and weaknesses of CAELD student writing; and

4. *Approach to marking* [AM]: perception of how the raters approached the marking of a CAELD essay.

See Appendix I for the rater interview protocol.

**Phase 2.** Three instruments were used in Phase 2: (a) teacher interview protocols, (b) classroom observation protocol, and (c) document analysis.

*Teacher Interview Protocol.*** For Phase 2, the teacher was interviewed four times over a period of four months. There were three main points of interest for the teacher interview protocol:

1. *Teacher’s background information* [TBI]: experience as an EAP teacher and experience using diagnostic assessment to inform classroom teaching;

2. *Decision-making process* [DM]: insight into the teacher’s decision-making process for implementing and using the CAELD feedback, and the tensions and benefits of using the CAELD for diagnostic purposes; and

---

⁴ The short form represents the points of interest in the interview protocols, and is used throughout this study.
3. *Teacher’s Perceptions [TP]:* teacher’s on-going perceptions of using the CAELD for informing classroom teaching, as well as the usefulness of the feedback on teaching and student learning.

See Appendix J for the three teacher interview protocols.

**Classroom Observation Protocol.** The classroom observation protocol was used to gather insight into how the teacher integrated the diagnostic feedback into her classroom teaching.

Direct observations allow for the researcher to gain a deep understanding of the context (Patton, 2002). In language testing, there has been a call in the literature for more studies that explore the classroom assessment process through observational data (Leung & Lewkowicz, 2006). The concern with observational data is that there is often too much data to record and distill later during the analysis stage (Patton, 2002). In order to focus on specific aspects of the setting, Cresswell (2007) recommends using an observation protocol. The classroom observation notebook used in this study was modelled on Hawkey’s (2006) work (see Appendix K). To aid in organizing the classroom observation protocol, the following note-taking devices were employed: a square box to indicate when the teacher wrote instructions on the board, a “T” (Teacher) and “S” (Student) followed by a colon to indicate the role of the speaker, and square brackets to indicate a description of the physical movement of the teacher and/or student. The focus of the protocol was on the specific implementation of the diagnostic feedback into classroom instruction. The purpose of this classroom observation was to supplement the information obtained from the two other data sources: interviews and document analysis.

**Document analysis.** Although document analysis was not an instrument, it is important to describe my rationale for including the method in this study alongside the interview and classroom observation protocols. Document analysis of the course syllabus, handouts, and
meetings provided a supplementary source of evidence of how the teacher incorporated the CAELD into her instruction (Bowen, 2009); the documents communicated her expectations, as well as instructions that were perhaps not included in the other sources of data.

**Phase 3.** Two instruments were used in Phase 3: (a) reflective tasks and (b) student interview protocol.

**Reflective tasks.** Three reflective tasks were given to the students for them to complete as part of their coursework. The purpose of the three reflective tasks was to understand the students’ perceptions of the diagnostic feedback and the influence of the feedback on their learning throughout the course. This method was similar to an informal questionnaire. The reflective tasks contained Likert-style, yes/no, and open-ended questions that asked the students to reflect on their experience with the diagnostic assessment feedback. There were two main points of interest in the reflective tasks: (a) perceived usefulness of diagnostic feedback for student learning and (b) influence of the diagnostic feedback on student learning. In addition, the reflections contained questions related to the appropriateness of the feedback and background information about the students. The questions that corresponded to the perceived usefulness were drawn from Jang’s (2005, 2009) work on students’ perceptions of diagnostic feedback. See Appendix L for the three reflective tasks.

**Student Interview Protocols.** For Phase 3, the students were interviewed twice. There were three main points of interest for the student interview protocol:

1. *Students’ Background information* [SBI]: background information about students’ home country, first language, age, previous language-learning experience, and previous experience with diagnostic feedback;
2. *Use of diagnostic feedback* [UDF]: insight into students’ use of the diagnostic assessment feedback to support learning in and out of the classroom, and elements that contribute to the successful use of the diagnostic feedback; and

3. *Students’ Perceptions* [SP]: students’ perceptions of and reactions to the diagnostic feedback, and perceptions of diagnostic assessment as a concept to help student learning and teaching.

There was an overlap between the questions asked in the student interview protocol and the reflections. This overlap allowed for the students to expand on their answers. I was also able to ask follow-up questions with the students, which I was not able to do with the reflective tasks (see Appendix M for the student interview protocols).

**Data Collection**

The study was carried out in three concurrent phases, from December 2009 to May 2010.

**Phase 1.** Two experienced CAEL raters were invited to re-mark 11 CAEL essays (22 essays total) at band scores representative of the CAEL band score for the intermediate-level course in the EAP program taught by the teacher in Phase 2. The band scores observed in Phase 2 ranged from 20-5 to 60. Thus there were two essays selected for each sublevel (e.g., 50 or 50-) (see Table 6). The sublevels represented borderline scores where an essay had characteristics of two levels. Included in the table is also the score assigned to the essay by the raters; essays were then rank ordered and assigned reference numbers based on the order. When referring to the actual essay or the student who wrote the essay, I use the essay code, e.g., essay 2 or student 2 corresponding to the number in the right-hand column of the table. The two raters completed the think aloud protocols while reviewing the 11 essays (Lazaraton & Taylor, 2007). Raters were

---

5 When scoring the CAEL essays, the raters will include symbols (- or +) to indicate whether the essay is borderline to the next lower or higher band score.
trained in person by me, the researcher, on the process of conducting the think aloud protocols.

However, the raters performed the think aloud protocols at home, so they could be at their leisure to complete the protocols. Catherine completed her set of verbal protocols in mid-February 2010 and Isobel in early March 2010. All of the think aloud protocols were transcribed verbatim. The references used for the rater think aloud in reporting the findings corresponded to the essay that was read by the rater, as seen in Table 6, e.g., Think Aloud 1, 2, 3.

Table 6

*Essays Included for the Think Aloud Protocols*

<table>
<thead>
<tr>
<th>Rater</th>
<th>Original Score</th>
<th>Think Aloud Score</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Catherine</td>
<td>40-</td>
<td>60</td>
<td>Essay 1</td>
</tr>
<tr>
<td>Catherine</td>
<td>60-</td>
<td>60</td>
<td>Essay 2</td>
</tr>
<tr>
<td>Catherine</td>
<td>50+</td>
<td>60</td>
<td>Essay 3</td>
</tr>
<tr>
<td>Catherine</td>
<td>50</td>
<td>50+</td>
<td>Essay 4</td>
</tr>
<tr>
<td>Catherine</td>
<td>40+</td>
<td>50</td>
<td>Essay 5</td>
</tr>
<tr>
<td>Isobel</td>
<td>60-</td>
<td>50</td>
<td>Essay 6</td>
</tr>
<tr>
<td>Isobel</td>
<td>50</td>
<td>50/50-</td>
<td>Essay 7</td>
</tr>
<tr>
<td>Catherine</td>
<td>30-</td>
<td>50-</td>
<td>Essay 8</td>
</tr>
<tr>
<td>Catherine</td>
<td>40</td>
<td>40+</td>
<td>Essay 9</td>
</tr>
<tr>
<td>Isobel</td>
<td>50+</td>
<td>40</td>
<td>Essay 10</td>
</tr>
<tr>
<td>Isobel</td>
<td>50-</td>
<td>40</td>
<td>Essay 11</td>
</tr>
<tr>
<td>Isobel</td>
<td>40+</td>
<td>40</td>
<td>Essay 12</td>
</tr>
<tr>
<td>Isobel</td>
<td>40-</td>
<td>40</td>
<td>Essay 13</td>
</tr>
<tr>
<td>Catherine</td>
<td>50-</td>
<td>40-</td>
<td>Essay 14</td>
</tr>
<tr>
<td>Catherine</td>
<td>30</td>
<td>40</td>
<td>Essay 15</td>
</tr>
<tr>
<td>Isobel</td>
<td>40</td>
<td>30+</td>
<td>Essay 16</td>
</tr>
<tr>
<td>Isobel</td>
<td>30-</td>
<td>30/30+</td>
<td>Essay 17</td>
</tr>
<tr>
<td>Catherine</td>
<td>30-</td>
<td>30</td>
<td>Essay 18</td>
</tr>
<tr>
<td>Catherine</td>
<td>30+</td>
<td>30</td>
<td>Essay 19</td>
</tr>
<tr>
<td>Isobel</td>
<td>30+</td>
<td>30</td>
<td>Essay 20</td>
</tr>
<tr>
<td>Isobel</td>
<td>30</td>
<td>30</td>
<td>Essay 21</td>
</tr>
<tr>
<td>Isobel</td>
<td>20+</td>
<td>20+</td>
<td>Essay 22</td>
</tr>
</tbody>
</table>
Two semistructured interviews were conducted with the raters within one week after they finished the think aloud protocols. The interviews with Catherine and Isobel took a total of 60 and 75 minutes, respectively, and were audio recorded and transcribed. The reference used for the rater interviews in the write up of the findings in Chapter 4 was: (Rater Pseudonym) Interview.

**Phase 2.** This phase explored how one EAP teacher integrated diagnostic feedback into her teaching. The instructionally supportive environment provided to the teacher offered the opportunity to track her decision-making process in an optimal classroom setting. Data collection strategies included classroom observations; document analysis of course syllabus, instructional materials, and biweekly meetings between the instructor and researcher; and semistructured interviews with the teacher to track the teacher’s decision-making process when using the diagnostic feedback for instruction. Table 7 provides a summary of the data collected for Phase 2.

The classroom observations demonstrated how the teacher integrated the feedback with her classroom teaching. Of particular interest were instances of the instructor giving the diagnostic feedback back to the class or to individual students. Of the 20 classroom observations conducted, five were selected for analysis. The five classroom observations chosen were based on key points in the classroom instruction that related to the diagnostic assessment. The individual student-teacher interactions were only audio recorded if the students had formally consented and agreed to be recorded. The field notes (reflective and descriptive), supported by the audio recordings, were used to form a comprehensive record of the classes observed. Portions of the five classroom observations chosen were transcribed.
Table 7

**Summary of Phase 2 Data Collected**

<table>
<thead>
<tr>
<th>Data Collection Method</th>
<th>Date Collected</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Semistructured Interviews</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interview 1</td>
<td>Jan. 21, 2010</td>
<td>Diane Interview 1</td>
</tr>
<tr>
<td>Interview 2</td>
<td>Feb. 8, 2010</td>
<td>Diane Interview 2</td>
</tr>
<tr>
<td>Interview 3</td>
<td>Apr. 27, 2010</td>
<td>Diane Interview 3</td>
</tr>
<tr>
<td><strong>Document Analysis</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meetings (5)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meeting 1: Preliminary meeting</td>
<td>Dec. 8, 2009</td>
<td>Meeting 1</td>
</tr>
<tr>
<td>Meeting 2: Looking through results</td>
<td>Jan. 13, 2010</td>
<td>Meeting 2</td>
</tr>
<tr>
<td>Meeting 3: Course planning</td>
<td>Jan. 26, 2010</td>
<td>Meeting 3</td>
</tr>
<tr>
<td>Meeting 4: Reflecting on student reactions to results</td>
<td>Jan. 28, 2010</td>
<td>Meeting 4</td>
</tr>
<tr>
<td>Meeting 5: Looking through reflections</td>
<td>Feb. 3, 2010</td>
<td>Meeting 5</td>
</tr>
<tr>
<td>Instructional Materials (6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Document 1: Course syllabus</td>
<td>Jan. 5, 2010</td>
<td>Document 1</td>
</tr>
<tr>
<td>Document 2: Portfolio criteria</td>
<td>Feb. 8, 2010</td>
<td>Document 2</td>
</tr>
<tr>
<td>Document 3: Example learner profile (Teacher)</td>
<td>Dec. 8, 2009</td>
<td>Document 3</td>
</tr>
<tr>
<td>Document 4: Example learner profile (Student)</td>
<td>Dec. 8, 2009</td>
<td>Document 4</td>
</tr>
<tr>
<td>Document 5: Sample speed reading exercise</td>
<td>Feb. 8, 2010</td>
<td>Document 5</td>
</tr>
<tr>
<td><strong>Classroom Observations (CO) (5)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CO 1: Administration of the CAEL (Section B)</td>
<td>Jan. 7, 2010</td>
<td>Classroom Observation 1</td>
</tr>
<tr>
<td>CO 2: Delivering of feedback to students (Section A)</td>
<td>Jan. 28, 2010</td>
<td>Classroom Observation 2</td>
</tr>
<tr>
<td>CO 3: Individual student-teacher interactions (Section B)</td>
<td>Jan. 29, 2010</td>
<td>Classroom Observation 3</td>
</tr>
<tr>
<td>CO 4: Speed reading instruction (Section A)</td>
<td>Feb. 8, 2010</td>
<td>Classroom Observation 4</td>
</tr>
<tr>
<td>CO 5: Small group work (Section B)</td>
<td>Mar. 13, 2010</td>
<td>Classroom Observation 5</td>
</tr>
</tbody>
</table>

Since I, the researcher, played a role in facilitating the implementation of the CAELD as an instructional tool, it is important to describe my role in the course. My role as a researcher was that of both an observer and a facilitator. As a facilitator, I helped support the teacher by
completing administrative tasks for her. In addition, through the meetings with the teacher, I saw myself as an active listener, asking for clarification to help further my understanding of the situation. This questioning would sometimes stimulate the teacher to further her thinking about the diagnostic assessment, the planning of tasks, and other decisions related to the course. The meetings ranged from 10 to 90 minutes; there were 10 meetings conducted, but only five were selected for analysis and transcribed. The five meetings were chosen based on whether or not the discussions expanded on the teacher’s instructional decisions seen in the classroom observations.

Three semistructured interviews took place with the teacher at different stages in the implementation of the diagnostic assessment feedback (beginning, middle, and end of the term). The interviews were up to 90 minutes long and were audio recorded. All of the interviews were transcribed.

**Phase 3.** There were two data collection techniques used in Phase 3: (a) reflective tasks given to all of the students and (b) two semistructured interviews with nine students. The three reflective tasks were collected from the students three times, at the beginning (January 27/28), middle (March 3/4), and end (March 22/23) of the term. The reflective tasks were similar to informal questionnaires because the instructor and I walked around the classroom and made sure that the students understood the questions being asked of them. In a setting where formal questionnaires are administered, students may not be able to ask questions about what is being asked of them. In the write-up of the findings, the reflections are referred to as (Student Number) Reflections 1, 2, or 3.

Nine students also participated in two semistructured interviews (at the beginning and end of the term). To randomly select the nine students, I took a stack of consent forms and selected the first 10 names. Six initially responded. I then emailed four more from the stack and
three responded positively. The first and second sets of interviews were conducted, respectively, between February 1 to 16 and March 22 to 31. The interviews were each 60 to 90 minutes in length and were audio recorded. In the write up of the findings, the interviews are referred to as (Student Name) Interview 1 or 2.

Data Analysis

I employed a three-cycle approach to analyze the three phases of data. I purposely adopted the term cycle (Saldaña, 2009) to convey the cyclical nature of the data analysis progression. The first cycle formed the initial coding and familiarization of the data for each phase. The second cycle considered the data in relation to the research questions and previous research for that phase; for example, Phase 1 (Scoring) was compared to research questions 4.1 to 4.3 and to the CAELD scoring guide. I conducted the first and second cycle analyses sequentially and independently for each phase. Four levels of labelling were adopted: subcodes, codes, subthemes, and themes (Saldaña). This process of analysing the data first inductively and second deductively is congruent with Patton’s (2002) description of modified analytic induction as being: “first deductive and then inductive as when, for example, the analyst begins by examining the data … by applying a theoretical framework” (p. 454). Another key feature of analytic induction is seeking out negative cases to provide rival explanations. This process of seeking explanations is in line with the argument-based approach to validation, which was carried out in the third cycle analysis. The third cycle included examining the case as a whole (the findings from the three phases) to arrive at an integrative evaluation of the validity of the CAELD with an emphasis on its use for informing classroom instruction. This third cycle analysis was based on the AUA for the CAELD. This level of analysis was done by examining the evidence, in light of the research questions from across the three phases, in order to locate
backing for the warrants to support the claims or to identify rebuttals that formed the counter
evidence for the warrants.

Three hermeneutic units were created in Atlas.ti, one for each phase. All of the transcripts
from the audio recordings of think aloud protocols, interviews, meetings, transcribed portions of
the classroom observations with notes, and instructional materials were entered into Atlas.ti.
Under the heading *Second cycle analysis*, I describe how the portions of the classroom
observations were selected. Selective data from the reflective tasks were entered into Excel. I
created binders for the three phases. The Phase 1 binder contained all of the essays scored by the
CAEL raters, identified by their reference number. The Phase 2 binder contained four sections:
teacher interviews, meetings, classroom observations, and documents. There were two Phase 3
binders. The first held the sets of student interviews, ordered by the name assigned to each
participant. The second Phase 3 binder contained the photocopies of the reflective tasks,
organized by participant identification number and by when the reflections were administered
(beginning, middle, or end of the course).

**First cycle analysis.** To become acquainted with the data and establish initial coding
frameworks, select transcripts were printed off and coded by hand. In Phase 1, I chose to print
off four think aloud transcripts (two essays, one high- and one low-scoring, from each of the two
raters). In Phase 2, I selected the three teacher interview transcripts to print off and read. For
Phase 3, I first looked at four interview transcripts (Douglas’ and Dennis’ interviews at the
beginning and end of the course). After I analyzed the initial student interviews, I looked through
the reflective tasks to identify the key questions across the three reflective tasks to enter into
Excel.
While reading the transcripts, I made comments in the margins regarding my preliminary observations of the data, which resulted in an initial code list. Once these select transcripts were read, the preliminary code list was entered into Atlas.ti. During this stage of the analysis, I wrote memos documenting my thoughts about the possible codes and making observations about the data. Memos are a well-established technique for analyzing qualitative data (e.g., Cresswell, 2007; Miles & Huberman, 1994). I wrote a memo in Atlas.ti anytime I noted that a new code or a modification to an existing code was needed. When coding the transcripts by hand, I wrote the memos using sticky notes when reading the transcripts and then entered them into Atlas.ti.

Second cycle analysis. This second cycle of the analysis served as a way to link the research questions, scoring guide (in Phase 1), and previous research to the data. There were five steps in the second cycle analysis: (a) considering the preliminary coding, from the first cycle analysis, in relation to research literature, (b) looking for correspondence between the data and the research questions, (c) revising the codes to be as meaningful as possible, (d) identifying patterns among the codes to form themes, and (e) applying the coding framework. In applying the coding framework to all types of data for the phase, the codes were expanded or refined to fit the new data. Each step is described in detail for the three phases.

Phase 1. In this second cycle, the coding for Phase 1 was slightly different from that for Phases 2 and 3. The first theme was actually established last, after all of the codes and other themes were established and refined due to the fact that I analyzed the think alouds in two ways. First, I coded all of the think alouds and entered them into Atlas.ti; these codes informed themes 2 and 3, as well as the parallel theme (see Table 8). Table 8 illustrates the codes identified in the first and second cycle coding for Phase 1. Second, I decided to analyze the think aloud data for potential rater effect based on the score discrepancies between the original and think aloud
scores. To do so, I analyzed the individual think alouds holistically to detect the raters’ reading strategies or styles, which formed the first theme of the *Rater decision-making process*. The codes informing this first theme originated from Smith’s (2000) research on rater behaviour.

Table 8

**Phase 1 Coding Framework**

<table>
<thead>
<tr>
<th>First Cycle Coding</th>
<th>Second Cycle Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Preliminary Code List)</td>
<td>(Revised Code List)</td>
</tr>
<tr>
<td><strong>Theme 1: Rater Decision-making process</strong></td>
<td><strong>Theme 2: Dimensions of Student Writing</strong></td>
</tr>
<tr>
<td>Code 1: Reading strategy</td>
<td>Code 3: Strengths</td>
</tr>
<tr>
<td>Code 2: Score decision</td>
<td>Code 4: Weaknesses</td>
</tr>
<tr>
<td><strong>Theme 2: Dimensions of Student Writing</strong></td>
<td>Code 5: Developing ability</td>
</tr>
<tr>
<td><strong>• Strengths</strong></td>
<td><strong>• Weaknesses</strong></td>
</tr>
<tr>
<td>Code 6: Learn and review</td>
<td>Code 7: Practice</td>
</tr>
<tr>
<td>Code 8: Tasks</td>
<td>Code 9: Strategies</td>
</tr>
<tr>
<td><strong>• Transitioning trait</strong></td>
<td><strong>• Grammar, Tense, Vocabulary, Clarity,</strong></td>
</tr>
<tr>
<td><strong>• Control, Argument, Cohesiveness, Linking Ideas, and Addressing the topic</strong></td>
<td><strong>• Code 10: Language</strong></td>
</tr>
<tr>
<td><strong>Parallel Theme: Features of Writing</strong></td>
<td><strong>• Code 11: Organization</strong></td>
</tr>
<tr>
<td><strong>• Content</strong></td>
<td><strong>• Code 12: Content</strong></td>
</tr>
</tbody>
</table>

Since there were one or sometimes two band score differences between the original scores and those assigned by the raters in the think alouds, I grouped the essays and think alouds into three levels of high, middle, and low to broaden the score distinctions from the original bandscores. The grouping of think alouds was almost evenly split between the high and middle levels. By grouping the essays into the levels, I was still able to look for trends across groups of
students with similar ability. The coding framework that emerged from the interview and think aloud data was analyzed in relation to this level grouping of high, middle, and low, as well as by rater.

Three changes were made to the preliminary code list to form the revised list for Phase 1 (see Table 8). First, I relabelled transitioning trait to developing ability based on Jang’s (2005) study. Second, I decided to include a parallel theme, entitled features of writing, which included the same quotations as the dimensions of student writing as well as the advice did. I realized it was necessary to understand if there were any differences in the strengths and weaknesses based on the criteria from the CAELD scoring guide. Third, as I expanded my coding to more think alouds during the second cycle, I recognized the need to have codes that described the types of advice being offered by the raters, rather than just the skill being targeted. As a result, I created subcodes drawing on the think alouds as well as the CAELD scoring guide (Appendix N).

Following the sequence of the coding framework, I wrote up the findings for Phase 1 in Chapter 4 and concluded the chapter by summarizing how the Phase 1 findings corresponded to the research questions.

Phase 2. In total, there were three types of transformations that occurred from the preliminary code list to the revised code list (Table 9). First, I combined multiple codes together to form a single code. For example, the initial codes of timing and organization were combined and re-coded into the new codes of planning and evaluation. In addition, I re-coded some quotations into different codes; these revised codes were more representative of the types of decisions that Diane seemed to make when using the CAELD in her classroom instruction, as seen in the classroom observations. The initial code of perceived success was added to the code student engagement to reflect the types of successful outcomes that Diane sought as confirmation
that the CAELD was useful. Similarly, the four initial codes of *making a connection for students, student centred feedback, self-directed group, and support to students* was add to *self-directed learning* to better encompass how Diane encouraged the students’ use of the CAELD feedback.

The second type of transformation made to the preliminary code list was renaming the initial codes. I renamed two initial codes *targeted groups* and *confidence using the CAELD*. The former was revised to *instructional support* to be more representative of the function that the targeted groups served Diane in her instruction and use of the CAELD feedback. The latter was renamed to *diagnostic competence*, drawing on Edelenbos and Kubanek-German’s (2004) term. Once the codes were combined and renamed, I looked for patterns across the codes to establish the themes for the final type of alteration made. The themes helped organize the codes into meaningful pieces (Patton, 2002). Because the coding process was iterative, the codes were also examined again for their fit within the established themes.

After the codes and themes were adjusted, I applied the codes to the remaining meeting transcripts in order to examine the viability of the framework. Five classroom observations were selected as an additional data source to which to apply the coding framework. I listened to the audio recordings of the classroom observations while also reading the corresponding observation protocols. Any time I noted an occurrence that was reflected in the coding framework, I would note the time and jot down the synopsis of the instance in the Phase 2 research journal. These time stamps were then transcribed, and the coding framework was applied to the documents collected.
### Phase 2 First and Second Cycle Coding Framework

<table>
<thead>
<tr>
<th>First Cycle Coding (Preliminary Code List)</th>
<th>Second Cycle Coding (Revised Code List)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Theme 1: Background and Context</strong></td>
<td></td>
</tr>
<tr>
<td>• Background</td>
<td>• Code 1: Background</td>
</tr>
<tr>
<td>• Course sections</td>
<td>• Code 2: Teaching context and course goals</td>
</tr>
<tr>
<td>• Background</td>
<td>• Code 3: Experience with diagnostic assessment</td>
</tr>
<tr>
<td>• Approach to assessment and teaching</td>
<td>• Code 4: Approach to assessment</td>
</tr>
<tr>
<td></td>
<td>• Code 5: Approach to teaching and learning</td>
</tr>
<tr>
<td>• Learning expectations</td>
<td>• Code 6: Learning expectations</td>
</tr>
<tr>
<td><strong>Theme 2: Decision-making process</strong></td>
<td></td>
</tr>
<tr>
<td>• Timing</td>
<td>• Code 8: Planning</td>
</tr>
<tr>
<td>Organization</td>
<td>• Code 9: Evaluation</td>
</tr>
<tr>
<td>• Skills and strategies</td>
<td>• Code 10: Skills and strategies</td>
</tr>
<tr>
<td>• Targeted groups</td>
<td>• Code 11: Instructional support</td>
</tr>
<tr>
<td><strong>Theme 3: Using the CAELD</strong></td>
<td></td>
</tr>
<tr>
<td>• Confidence using the CAELD</td>
<td>• Code 12: Diagnostic competence</td>
</tr>
<tr>
<td>• The CAELD as a diagnostic assessment</td>
<td>• Code 13: The CAELD as a diagnostic assessment</td>
</tr>
<tr>
<td>Match between test and target content</td>
<td></td>
</tr>
<tr>
<td><strong>Theme 4: Spirit of Formative Assessment</strong></td>
<td></td>
</tr>
<tr>
<td>• Perceived success/benefit for students</td>
<td>• Code 14: Student perceptions</td>
</tr>
<tr>
<td>Student perceptions</td>
<td></td>
</tr>
<tr>
<td>• Student engagement</td>
<td>• Code 15: Student engagement</td>
</tr>
<tr>
<td>• Student centered feedback</td>
<td>• Code 16: Self-directed learning</td>
</tr>
<tr>
<td>Self-directed group</td>
<td></td>
</tr>
<tr>
<td>Making a connection for students</td>
<td></td>
</tr>
<tr>
<td>Support to students</td>
<td></td>
</tr>
<tr>
<td>• Confirmation of teaching</td>
<td>• Code 17: Refined teaching practice</td>
</tr>
<tr>
<td>Teaching process</td>
<td></td>
</tr>
<tr>
<td>• Change in teaching/influence on future</td>
<td>• Code 18: Influence on future teaching</td>
</tr>
<tr>
<td>teaching</td>
<td></td>
</tr>
<tr>
<td>• Feedback</td>
<td>• Code 19: Formative feedback</td>
</tr>
</tbody>
</table>
Phase 3. Similar to the other two phases, there were changes from the initial coding in the first cycle analysis to the refined code list developed in the second cycle analysis (Table 10). First, I combined two codes into one to be more representative of the overlap occurring in the initial codes. These codes, *perceptions of Diane’s teaching* and *perceptions of targeted groups*, were unified into a single code, which represents the teaching that the students encountered, whether by Diane or by the teaching assistants. In addition, the two codes of *learning goals* and *learning styles* were collapsed into the single code of learning goals. I then revised the code labels for: (a) *reactions to results*, (b) *strategies to target learning*, and (c) *conceptions of diagnostic assessment*. The word *use* was added to the first, to incorporate the reflection data about students’ reported thinking about the CAELD feedback inside and outside of class.

Second, the word *approaches* replaced *strategies*, in order to be more inclusive of the other types of activities in which the students engaged for targeting their weaknesses. The third revision was made to include the word *evolving* when it became apparent that the students’ ideas of diagnostic assessment changed during the intermediate-level EAP course.

Once it became apparent that the students’ conceptions of diagnostic assessment might have changed from the first interview to the second, I created a table that listed the quotations side-by-side to further analyze this possible change. Subcodes of *formative* and *evaluative* were used to note these changes. The final type of alteration to the preliminary code list resulted from organizing the codes into themes. The three themes identified were based on the different stages the students seemed to experience: the CAELD, the instruction based on the feedback, and the use of the feedback for their own learning. After noticing the changes in the students’ definitions of diagnostic assessment I realized that the *Evolving conceptions of diagnostic assessment* was more illustrative of the students’ learning than of their test-taking experience, which led to
grouping the code within *Targeted Learning*. Please see Table 10 for the initial and revised coding framework for Phase 3. While I have sequentially described the changes made during the second cycle analysis, the changes were iterative and non-linear. The findings for Phase 3 appear in Chapter 6.

Table 10

*Phase 3 Coding Framework*

<table>
<thead>
<tr>
<th>First Cycle Coding</th>
<th>Second Cycle Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preliminary Code List</td>
<td>Revised Code List</td>
</tr>
</tbody>
</table>

Theme 1: *Experiencing the CAELD*
- Administration
- Perceptions of the CAEL

Theme 2: *Experiencing Diagnostic Teaching*
- Reactions to the results
- Perceptions of Diane’s teaching
- Perceptions of targeted groups
- The reflections

Theme 3: *Targeted Learning*
- Reactions to the results
- Strengths and weaknesses
- Conceptions of diagnostic assessment
- Learning goals
- Learning styles
- Strategies to target learning

Third cycle analysis. The goal of the third cycle analysis was to arrive at a coherent understanding of the case in its entirety—the use of the CAELD within an EAP program at a Canadian university. This final cycle of analysis unified the direct observations and interpretations of the three stakeholder perceptions into an integrated description of validating the CAELD based on the AUA specified in Chapter 2. To achieve this goal, I mapped out the
findings from the three phases (embedded units of analysis) to the four assessment claims, using Toulmin’s (2003) logic of warrants and backing. This formal, third cycle analysis began once the findings from the three results chapters were written up.

An informal analysis of the concepts and the relationship to validity occurred when I made notes of how my interpretations and observations connected to validity theory. This continual consideration of how the data related to validity theory was expected, given that the research questions were derived directly from the AUA validity framework. These thoughts were recorded in a research journal that was used throughout the data analysis and write-up process of this study. In writing up the third cycle analysis, I reported the case level findings, organized by the assessment claims: assessment claims, interpretations, decisions, and consequences (Bachman & Palmer, 2010).

Criteria for Judging Qualitative Evidence

This study incorporated a variety of strategies to enhance the integrity of the research by drawing on Lincoln and Guba’s (1985) criteria for judging qualitative evidence, particularly (a) credibility and trustworthiness, (b) confirmability and dependability, (c) authenticity and reflexivity, and (d) transferability. Credibility and trustworthiness speaks to the overall design of the study and its purpose: to gain a comprehensive understanding of the “truth” about the phenomenon or perspective being studied. In this study, multiple sources of data were collected for each phase, to provide triangulation of the findings or possible counter evidence. Further, I immersed myself in the research context—particularly, the classroom—to acquire an in-depth appreciation of the dynamic interactions between the teacher and students.

Confirmability and dependability refer to the integrity of the data analysis and interpretation of the findings. To document my thoughts and interpretation of the findings, a
research notebook was kept throughout the entire process, beginning prior to data collection and continuing through to the final write-up. The use of memoing throughout the coding process provided an additional means for taking stock of any changes in my interpretations of the data.

Authenticity and reflexivity looks to the positioning of the researcher to the study and her potential biases. In Chapter 1, I locate myself in relation to the study by providing a description of my background and interest in the phenomenon of validating a large-scale test for the additional purpose of diagnostic assessment in the classroom. Further, the act of keeping a research journal is a reflexive practice when re-reading entries to look for possible sources of bias.

Transferability considers the positioning of the study within the broader field. In other words, it considers the relevance of the findings to other contexts. The reporting of the findings are rich in detail, providing a thick description (Geertz, 1973) of the context. The purpose of this attention to detail is to enable readers to make their own interpretations about the findings’ applicability to other settings, and to conduct similar research. Such research could then lead to generalization of the findings and/or expansion of the theoretical understanding of how diagnostic assessment is used and validated. Transferability in a case study approach is gained through the interpretation of the findings to existing theoretical understanding of the phenomenon (Yin, 2009).

Summary

This case study examined the use of the CAELD in an EAP program from the rater, teacher, and student perspectives. To accomplish this aim, multiple sources of data were gathered as evidence to support and possibly refute the assessment claims articulated in Chapter 2. The
next three chapters on scoring, teaching, and learning report on the findings from these three perspectives.
Chapter 4: Scoring

The assessment is a snapshot of a test takers’ skill set to predict how likely they’re to perform in an academic setting, that’s what we’re doing to place them. So it needs to indicate in some way where the test taker needs more help in language skill development for academic study. We need to understand that language skill development is a process. So, we need to assess weak areas and find the best starting point for skill improvement. (Isobel Interview)

Overview

The above quote by Isobel, one of the two raters interviewed for Phase 1 of the study, captures the essence of this chapter: to investigate the starting point for targeting students’ needs in writing based on the CAELD essay with meaningful advice on how to move forward. In other words, this chapter reports on the findings of Phase 1 from the raters’ think alouds and interviews. The think alouds recorded the raters’ perceptions of the diagnostic subcomponents in the CAELD essays and the advice raters would later offer to the students. The interviews elaborated on the raters’ approach to scoring the CAELD essays and their perceptions of it as a diagnostic tool. This chapter is organized into three parts. First, I review the research question and facilitating questions guiding this phase of the study and outline the themes and codes. Second, I report on the findings, organized by the themes. Finally, I conclude this chapter with a discussion of the findings in relation to the facilitating questions and research question considered.

Research Question

The following research question was addressed in Phase 1:

1. What are the features of CAELD scoring that enable it to be used for diagnostic purposes?
This research question was further expanded through the following facilitating questions:

1.1. How did the raters pinpoint the diagnostic subcomponents about the students’ writing ability?

1.2. What were the dimensions of the diagnostic subcomponents as observed by the raters?

1.3. What advice did the raters offer to the students on how to improve their academic writing skills?

**Themes and Codes**

There were three main themes and one parallel theme for Phase 1. The parallel theme (features of writing) included more quotations than any of the other themes. In fact, there were more quotations coded for the features of writing parallel theme than for Themes 2 (dimensions of student writing ability) and 3 (advice offered) combined (Table 11). The parallel theme served as an additional means to interpret the data, and is discussed in the three sections describing the three main themes. The parallel theme included 15 subcodes and was coded at a finer grain level. For example, a quotation coded as strength might have included two features of writing. Please see Appendix N for a full description of the subcodes informing the language, organization, and content codes.

**Raters’ Decision-Making Process**

To illustrate the raters’ decision-making process, I first outline the tendencies exhibited by the two raters, along with the features of writing they focused on when identifying the students’ strengths and weaknesses. Second, I describe the raters’ individual rating style, basing my observations on two codes: Reading Strategy and Perceptions of the CAELD essay.
Table 11

*Frequency Counts of Phase 1 Themes and Codes*

<table>
<thead>
<tr>
<th>Phase 1 Themes and Codes</th>
<th>Total Frequency (330)¹</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Theme 1: Raters’ Decision-Making Process</strong></td>
<td></td>
</tr>
<tr>
<td>Code 1: Reading Strategy</td>
<td>22</td>
</tr>
<tr>
<td>Code 2: Perceptions of the CAELD Essay</td>
<td>14</td>
</tr>
<tr>
<td><strong>Theme 2: Dimension of Student Writing Ability</strong></td>
<td>222</td>
</tr>
<tr>
<td>Code 3: Strengths</td>
<td>55</td>
</tr>
<tr>
<td>Code 4: Weaknesses</td>
<td>152</td>
</tr>
<tr>
<td>Code 5: Developing Ability</td>
<td>15</td>
</tr>
<tr>
<td><strong>Theme 3: Advice Offered</strong></td>
<td>72</td>
</tr>
<tr>
<td>Code 6: Review and Learn</td>
<td>12</td>
</tr>
<tr>
<td>Code 7: Practice</td>
<td>33</td>
</tr>
<tr>
<td>Code 8: Tasks</td>
<td>16</td>
</tr>
<tr>
<td>Code 9: Strategies</td>
<td>11</td>
</tr>
<tr>
<td><strong>Parallel Theme: Features of Writing</strong></td>
<td>435</td>
</tr>
<tr>
<td>Code 10: Language</td>
<td>227</td>
</tr>
<tr>
<td>Code 11: Organization</td>
<td>126</td>
</tr>
<tr>
<td>Code 12: Content</td>
<td>82</td>
</tr>
</tbody>
</table>

*Note.* The total frequency count does not include the parallel theme since the number of quotations coded are also included in Themes 2 and 3. Some of the quotations included in Themes 2 and 3 were coded more than once in the parallel theme.

When the band score differences were marked as an increase or decrease, there was a trend in the raters’ behaviour (Table 12). Catherine raised the band scores on five essays, kept
five the same, and lowered one; in contrast, Isobel did not raise any of the scores, kept seven the same, and lowered four. This trend suggests that Catherine is a lenient rater while Isobel is a severe rater. The rater preference to be lenient or severe was limited to essays grouped into the middle- and high-levels. That the inconsistencies in marking occurred only within certain proficiency levels demonstrates that Catherine and Isobel generally agreed on the quality of the writing in the lower level essays, but diverged for the higher level essays, where the students’ writing begins to differ more in terms of language, organization, and content features. These findings are supported by the research literature. Schaefer (2008) found that raters were more likely to differ with higher ability writers, but less likely towards lower ability writers. However, he also noted that in the case of extremely low-ability writers, raters showed tendencies to differ how they assigned scores. In the current study, the raters’ score differences for the low-level essays were limited and, thereby, demonstrated less bias, suggesting that perhaps the writing was not as low-level as the lowest level writing samples seen in Schaefer’s (2008) study.

Isobel and Catherine differed in regards to which subcomponents of the features of writing (i.e., language, organization, or content) they focused on (Table 13). The subcomponents from Isobel’s think alouds were balanced between the language and organization features of the essays while Catherine’s think alouds centred on language-use subcomponents. In fact, Catherine identified more than twice as many language-related subcomponents as organization ones. The lowest category of subcomponents identified by both raters related to the content features of the CAELD essays.
Table 12

**Difference Between Original and Think Aloud (New) Score**

<table>
<thead>
<tr>
<th>Essay Level</th>
<th>Reference</th>
<th>Rater</th>
<th>Difference between Old and New Score Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Essay 1</td>
<td>Catherine</td>
<td>+ (^1)</td>
</tr>
<tr>
<td>High</td>
<td>Essay 3</td>
<td>Catherine</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Essay 8</td>
<td>Catherine</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Essay 5</td>
<td>Catherine</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Essay 4</td>
<td>Catherine</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>Essay 2</td>
<td>Catherine</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>Essay 7</td>
<td>Isobel</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>Essay 6</td>
<td>Isobel</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Essay 15</td>
<td>Catherine</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>Essay 9</td>
<td>Catherine</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>Essay 12</td>
<td>Isobel</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>Essay 13</td>
<td>Isobel</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>Essay 10</td>
<td>Isobel</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Essay 11</td>
<td>Isobel</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Essay 14</td>
<td>Catherine</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Essay 16</td>
<td>Isobel</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Essay 17</td>
<td>Isobel</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>Essay 18</td>
<td>Catherine</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>Essay 19</td>
<td>Catherine</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>Essay 20</td>
<td>Isobel</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>Essay 21</td>
<td>Isobel</td>
<td>/</td>
</tr>
<tr>
<td></td>
<td>Essay 22</td>
<td>Isobel</td>
<td>/</td>
</tr>
</tbody>
</table>

\(^1\)The three symbols indicate a band score increase (+), decrease (−), or no change (/).

It is not surprising that Isobel and Catherine focused on different features of the students’ writing; other researchers have observed a similar trend when examining raters’ use of a holistic scale to score writing. Barkaoui (2010), Smith (2000), and Wigglesworth (1993) noted that raters...
using a holistic scale will consistently arrive at the same score, but draw on different criteria in a scale to do so. Specifically, Wigglesworth (1993) found that some of the raters were severe towards grammar criteria but were lenient on fluency, while the reverse was observed for other raters. Thus there does not seem to be a consistent pattern for severe and lenient raters and the criteria they focus on; rather, there are individual differences across raters. Next, I describe Isobel’s and Catherine’s personal rating style.

Table 13

*Features of Writing Ability Discussed by the Raters During the Think Alouds*

<table>
<thead>
<tr>
<th>Type of subcomponent</th>
<th>Isobel</th>
<th>Catherine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>35%(45)</td>
<td>58%(179)</td>
</tr>
<tr>
<td>Language</td>
<td>39%(50)</td>
<td>24%(75)</td>
</tr>
<tr>
<td>Organization</td>
<td>26%(33)</td>
<td>17%(53)</td>
</tr>
<tr>
<td>Total</td>
<td>128</td>
<td>307</td>
</tr>
</tbody>
</table>

*Percentages were calculated to compare the writing features discussed by Isobel and Catherine.*

**Isobel.** During the think alouds, Isobel would read the essays in their entirety and then look back at the text to find evidence to support her scoring decision based on the CAELD performance criteria. Based on Smith’s (2000) classification of raters’ reading strategies, Isobel adopted a “read-through-once-then-scan” approach. Smith characterized this type of reading strategy as continuous reading without a break in the flow and subsequent scanning of the text for evidence. During her interview, Isobel alluded to the fact that she would sometimes even re-read an essay when marking CAEL essays; “as raters, we find stronger logic as we read it through a few times. You know we always miss something the first time” (Isobel Interview). During the interview, Isobel acknowledged the importance that teaching experience had in
scoring an essay, as it allowed her to compare the writing against students she had taught.

Though Isobel stated that she did not consciously alter her marking style of the CAELD essays, she remarked that she was shocked at the diverse range of writing from an intermediate-level EAP course (Phases 2 and 3), which might have influenced her scoring, “I think I was kind of stymied at the outset [because of the wide range of writing ability] when I just said all I can do is mark them” (Isobel Interview).

Smith (2000) noted that the two raters in his study who shared Isobel’s reading style had an overall preference for language features of writing, specifically, one had a preference for grammar and the other for punctuation. In contrast, the comments from Isobel’s interview suggest that she prioritized organization and content over language features. She felt that it was essential for the students to “display some comprehension of the source material … [by providing] information relevant … and supportive of the thesis” (Isobel Interview). She elaborated further by stating, “the strength of this ability, in my mind, is how relevant the point is and how well it is developed” (Isobel Interview). Isobel reflected that, if the writing followed the “basic conventions of writing, then [the student] can signify when a paragraph starts and ends” (Isobel Interview). For her, the greatest type of weakness was an inability to understand the prompt. These comments from her interview contradict the balanced approach she took in the think alouds to identify the subcomponents. Even though Isobel was relatively balanced in how she drew on the CAEL scoring guide (Table 13), the subcomponents focused primarily on vocabulary (19), clarity (12), or morphology (7), with the remainder dispersed across the other subcodes. This emphasis on vocabulary is connected to the students’ understanding of the content because Isobel would comment if they had not understood a term from the text.

When asked about language features such as grammar, Isobel remarked that CAEL or
CAELD raters “typically treat grammar errors at the level of communication.” She elaborated with “I mean we know that native speakers make grammar errors all the time, so the focus is not on grammar per se, in my mind, the focus is on communication” (Isobel Interview). The emphasis on logic and forming an argument was further articulated in the advice she offered to the students. Isobel often suggested that the students start with “a plan of their ideas, and outline examples for each point” (Think Aloud 11). Thus Isobel felt the most important skills students needed to successfully communicate in academic writing were, first, coherence and, second, content.

Isobel defined diagnostic assessment as a placement test to “slot [the students] to some extent” (Isobel Interview). She elaborated, stating that the purpose was really to determine the students’ “readiness for the course goal” (Isobel Interview). However, she felt the ultimate goal of diagnostic assessment was to help the teachers develop materials corresponding to the diagnostic feedback. To support students’ learning, she recommended that teachers start with the students’ strengths over weaknesses, in order to increase the students’ motivation. The quote at the beginning of this chapter speaks to Isobel’s viewpoint that language learning is a process. By alluding to the need to focus on students’ strengths as an important part of diagnostic assessment, she seems to speak to the need to remind students about how far they have come in their language-learning journey.

**Catherine.** At various points when reading an essay aloud, Catherine would interject comments about the essay’s features. She verbalized her initial reactions to the writing, noting whether the text included visible paragraphs or commenting about the quality of the first couple of sentences. Based on Smith’s (2000) study, Catherine exhibited a “first-impression-dominates” approach to reading the essays. Smith defined this approach by the way the raters read through
the essays and by the way they commented (while reading or at the end) on their impression of the essay overall, before scanning the essay for evidence to support their decision. The raters in Smith’s study who used this approach demonstrated a preference for coherence features. In contrast to Isobel’s think alouds, Catherine’s description of the students’ subcomponents were considerably more detailed; further, she highlighted strengths and weaknesses that she saw in the writing, but which were not included in the CAEL scoring guide.

In her interview, Catherine remarked that the major problems L2 students faced when entering university were language mistakes that often impeded understanding of the students’ writing. Catherine’s rationale for focusing on language issues in CAELD writing stemmed from her viewpoint that the students needed to be prepared for the realities of university and life. She gave the example of how her brother and father would not accept a résumé with grammar errors in it. This preference for language as the most important feature of writing seemed to frame how Catherine viewed the students’ strengths and weaknesses in writing (see Table 13). In fact, she commented that weaknesses in organization of essays or using the source material in the CAELD were easier to remedy for some students than for those who had language issues:

Hooray, if it’s just something as simple as organization or just saying to the student you have to use the content, [he/she] are actually capable of doing it. That’s really simple to fix. It’s the language that’s the problem, unfortunately. (Catherine Interview)

Catherine went on to state that content problems were easy to remedy at any proficiency level. In contrast, she felt that language problems were harder to change since students with a relatively high level of proficiency “can communicate the message fairly easily, but not well enough that [the reader does not] have to think about it” (Catherine Interview). She further
explained that the challenges higher proficiency students faced related to the tendency for L2 students’ language to fossilize and not improve in certain areas such as word order.

Catherine also referred to diagnostic assessment as “placement of students in a class” (Catherine Interview), and elaborated by describing how she would use such a tool in her own teaching. Catherine exclaimed that, as a teacher, she would focus on “major problems” (Catherine Interview), recognizing the limited number of teaching hours in a 12-week EAP course, combined with the challenge of meeting with students individually. These major problems reflected not only the features of writing that Catherine deemed as most important, such as grammar, but also the major weakness that kept students from effectively communicating their ideas.

Isobel’s and Catherine’s previous experiences teaching in the credit and non-credit program, along with their knowledge of scoring the CAEL essay, influenced how they approached the think alouds and, in turn, identified diagnostic subcomponents. While both raters emphasized the importance of communication as the central goal for students’ writing ability, Isobel prioritized logic and coherence, while Catherine focused on grammar and language use. They also differed in their reading style. Isobel adopted a read-through-once-then-scan approach and Catherine, a first-impression-dominates approach. These differences might have contributed to the scoring patterns that Isobel and Catherine exhibited (as severe and lenient raters, respectively). Nonetheless, their unique conceptualization of the CAELD essays allowed for a richer understanding of the strengths, weaknesses, and developing ability possible. Next, I characterize these dimensions of student-writing ability as described by the raters.
Dimensions of Student Writing Ability

Three dimensions emerged from the think aloud data to form a continuum, portraying the students’ abilities from weak to strong: weaknesses, developing ability, and strengths. While each of these dimensions had unique characteristics, there were common elements as well. Across the dimensions, the frequency and detail in the raters’ comments increased with the students’ ability level. For example, Catherine’s comments about the higher level essays increased in the depth and insight provided about the students’ weaknesses. At the beginning of Essay 2, a high-level essay, Catherine described the student’s challenges with a sentence in the introduction: “Lots of little errors here. Strange wording, he’s missing a verb, preposition problem” (Think Aloud 2). In this instance, Catherine highlighted three concrete challenges that she had in reading a sentence. She went on to make numerous comments about the essay’s strengths and weaknesses. In contrast, Catherine made very few remarks about Essay 18, a low-level essay, and most of those remarks were grounded in the CAELD scale descriptors. For example, Catherine’s remark, “in and out of focus, well, his focus is a bit wrong all the way through” (Think Aloud 18) echoes the “weak focus” criteria descriptor for the organization category of the CAEL scoring guide. Catherine was able to identify weaknesses for Essay 18, but not with as much depth as she did for Essay 2.

Table 14 demonstrates how the frequency of comments regarding weaknesses, strengths, and developing ability increased with the students’ writing ability. The number of weaknesses doubled in frequency for every level, whereas the number of strengths identified from the low to the mid-level essays increased seven-fold. To describe this continuum in more detail, I contrast the two bookends, weaknesses and strengths, and then characterize the centre, the developing ability.
Table 14

*Frequency of Dimensions of Student Writing Ability by Level of Essay*

<table>
<thead>
<tr>
<th>Level of Essay</th>
<th>Weaknesses</th>
<th>Strengths</th>
<th>Developing Ability</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>95</td>
<td>39</td>
<td>5</td>
<td>139</td>
</tr>
<tr>
<td>Middle</td>
<td>39</td>
<td>14</td>
<td>8</td>
<td>61</td>
</tr>
<tr>
<td>Low</td>
<td>18</td>
<td>2</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>Total</td>
<td>152</td>
<td>55</td>
<td>15</td>
<td>222</td>
</tr>
</tbody>
</table>

**Weaknesses.** A weakness was noted whenever the writing interfered with the raters’ comprehension of the text, or when there were minor errors. Thus there were two types of weaknesses: major and minor. The major weaknesses were often, in fact, a group of weaknesses, which combined to result in incomprehensible text. In such instances, both raters would need to re-read or slow down in order to understand what the student was trying to convey. Even when slowing down to re-read, the raters were still unable to interpret the student’s meaning. This was the case with Essay 19. Catherine slowed down to re-read the following: “In the other hand, we need to control the use of pesticide in the farms to the need only” (Think Aloud 19). She subsequently stated “I don’t know what that means” (Think Aloud 19). Catherine further stated that the student “didn’t even know a pretty common phrase, on the other hand” and had incorrect word order, subject-verb agreement, and inappropriate vocabulary. Minor weaknesses were mistakes that annoyed the raters but did not interfere with comprehension. For instance, in Essay 9, Catherine noted that the student had spelled *affect* instead of *effect*, but indicated that spelling was not something that was included in the scoring guide. She further remarked the mistake did not interfere with overall coherence of the essay, but that addressing the spelling issue might lead to a better grade on a paper for a university course.
At the higher level of essays, the raters were able to better articulate the students’ weaknesses. In addition, there were fewer weaknesses identified that contributed to the raters’ loss of comprehension of the text. Most were minor and related to subject-verb agreement, use of plurals, spelling, redundant vocabulary, and flexibility of language. The major weaknesses identified corresponded to control of complex or simple sentence structures, morphology, inappropriate punctuation, and inaccurate vocabulary. The weaknesses for the mid-level essays related to the three broad features of writing (e.g., language, organization, and content). For instance, Isobel noted that Student 13 was struggling with appropriate vocabulary and word endings: “[Isobel reading] products will be very harmer [end of reading], so the student needs to learn endings” (Think Aloud 13). As previously discussed, only a few weaknesses were discussed for the low-level essays. Most of the raters’ comments were very general and remarked on overarching concerns such as text that was not predictable, language constraint problems, or lack of clear organization.

**Strengths.** A strength is when the raters would comment on an aspect of the writing regardless if it helped push an essay into the next bandscore. A key characteristic of strengths was that they contributed to the raters’ comprehension of the text a contrast to the weaknesses that interfered with understanding. Moreover, these strengths were grounded in the raters’ comparisons of the writing to other text within the same essay or to other essays. Indeed, if a sentence was significantly better than the rest of the essay, the rater questioned whether or not the student had plagiarized. For Think Aloud 4, Catherine scanned the readings to check if the student had, in fact, copied a sentence that was markedly better than the other writing in the essay. She was pleased to note that the student had “used some of the same words, but [it was] not plagiarism: he’s taken it and he’s managed to apply it correctly” (Think Aloud 4). Based on
her comments, the passage was a portrayal of the student’s ability to write the content in his own words, using an accurate and complex sentence structure. The example demonstrates the student’s competence to understand source information well enough to modify the word order and vocabulary to make it his, as well as the student’s ability to write a sentence with several clauses, using the correct word order and tense. In other words, this strength was a compilation of multiple simple strengths.

The majority of the strengths identified for the high- and mid-level essays were a compilation of more than one simple strength. For the high-level essays, most of the strengths related to the language category of the scoring guide, primarily focusing on syntactic accuracy for both complex and simple sentence structures. Since the bulk of the essays scored in the high-level group were by Catherine (5 out of 7), this emphasis on language as a strength is not a surprise. Her comments centred on whether the essays had accurate subject-verb agreement, proper punctuation, appropriate phrasal verb combinations, or control of simple sentence structures. For the mid-level essays, the majority of the strengths identified focused on organizational features, which is no surprise since Isobel predominantly scored the mid-level essays. For instance, Isobel stated Student 13 “doesn’t have problems with language, she has problems with logic.” The student’s essay contained a number of syntactically correct simple sentences, but the connection between the sentences was not obvious, at least not to Isobel. Catherine also noted organizational features for mid-level essays. According to Catherine, Essay 19 had a long introduction, despite having little else written, “so, he’s got one, two, three, four, five, six, seven, eight, nine lines of an introduction, which does include a thesis statement and preview” (Think Aloud 19). By counting the number of lines that the student had written, Catherine emphasized the amount of text the student was able to produce. Catherine further
added that the student had some of the necessary elements of a thesis statement and an overview of the essay. Once again, the strengths for the high- and mid-level essays were often a combination of more than one strength, which resulted in a text that was comprehensible to the rater.

In contrast to the mid-and high-level essays, there were only two strengths identified for the low-level essays. Both corresponded to the fact that the student had “included a thesis statement connected to the prompt” as stated by Isobel (Think Aloud 20). The students who wrote these essays struggled to link their ideas together or use correct word order, even if the thesis statement was grammatically correct. When other simple strengths were viewed in context of the surrounding text, a more complex picture of the students’ writing ability was revealed. It is not surprising that the simple strength does not reflect the complete picture of students’ writing or possible next steps—it is only one characteristic of the students’ writing.

Developing ability. In between strengths and weaknesses in academic writing is developing ability. This dimension is realized when the student shows signs of improvement, thus demonstrating some understanding of one subcomponent, but is still developing in another subcomponent or in his or her application of it. For instance, Catherine reflected on a complex sentence in Essay 9 that had “plenty of errors in it in terms of word form,” but she acknowledged that the student “actually got the form of the complex sentence” (Think Aloud 9). Based on Catherine’s comments, the student needed to improve his understanding of the noun verb form of the word. However, he seemed to understand how to pair more than one clause together and form a complex sentence.

Developing ability was also identified when the rater noticed that the student was able to apply a subcomponent, but only sporadically. This was the case with Essay 19: Catherine
remarked the student “got the proper structure here, he’s got the form of the word correct, so that’s good. That’s better than previously in this essay” (Think Aloud 19). With this example, it seemed that, although the student had used incorrect forms for nouns and verbs in one part of the essay, he did not make the same mistake later on. This type of information about a student’s burgeoning skill would be useful to convey to teachers and students, so the student is able to apply it in a consistently correct way.

The appearance of a student’s effort to try a subcomponent was also identified as developing ability. Catherine remarked that she could tell the student who wrote Essay 5 was trying: "He's working for the structures. He's trying the passive form. He's not getting it right, but he's trying . . . at least he is trying to stretch his language ability" (Think Aloud 5). However, she elaborated, “I think this is probably; maybe this is as good as this person can do” (Think Aloud 5). According to Catherine, the student had the knowledge of the sentence structures and passive form, but his application was incorrect. If there is indeed an effort on the part of students to push beyond their current skill set and stretch their language ability, it may be the ideal subcomponent to target for feedback. The student may already be motivated to improve and with the appropriate guidance can.

While the dimensions formed a continuum from weak to strong (represented by weaknesses, developing abilities, and strengths), the trajectory was not linear. Some of the weaknesses were actually a compilation of multiple weaknesses, which led to a breakdown in the raters’ understanding. However, others were minor that did not interfere with the raters’ comprehension of the text. At the centre were the developing abilities. This dimension had characteristics of both weaknesses and strengths—an indication that the student was learning and trying to improve her or his writing ability. The strengths contributed to the raters’
comprehension of the text and were seen as the positive aspects of the writing. The challenge was to provide meaningful feedback to students on all three dimensions, along with advice on how to improve, and what to target first. The final theme discussed considers the raters’ advice.

Table 15

*Types of Advice Offered by the Raters*

<table>
<thead>
<tr>
<th>Review and Learn</th>
<th>Practice</th>
<th>Strategies</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Elements of basic or complex sentences (G)¹</td>
<td>• Writing complex sentences (G)</td>
<td>• Shorten introduction (S)</td>
<td>• Read and answer questions about the text (K)</td>
</tr>
<tr>
<td>• Word order (G)</td>
<td>• Subject-verb agreement (T)</td>
<td>• Spell keywords correctly (V)</td>
<td>• Write long lists and practice comma placement (G)</td>
</tr>
<tr>
<td>• Punctuation (G)</td>
<td>• Passive form (T)</td>
<td>• Check punctuation (G)</td>
<td>• Make a plan/list to organize thoughts (C)</td>
</tr>
<tr>
<td>• Articles (G)</td>
<td>• Sentence boundaries/run-on sentences (G)</td>
<td>• Check tense (T)</td>
<td>• Explain thoughts to help provide details (C)</td>
</tr>
<tr>
<td>• Plurals (V)</td>
<td>• Preposition use (V)</td>
<td>• Ask the basic questions of why, how, etc. to develop an argument (A)</td>
<td>• Practice editing tasks (G)</td>
</tr>
<tr>
<td>• Differences between similar-looking words (V)</td>
<td>• Verb versus noun forms (V)</td>
<td>• Identify information to use in support of essay arguments (Su)</td>
<td>• Comb subordinate and main clauses (G)</td>
</tr>
<tr>
<td>• Paragraphing (S)</td>
<td>• Identification of vocabulary in context (V)</td>
<td>• Identify parts of sentences (G)</td>
<td>• Identify prefixes and suffixes (T)</td>
</tr>
<tr>
<td>• Spelling (V)</td>
<td>•</td>
<td>• Learn to differentiate active and passive form (T)</td>
<td>• Make a chart on parts of speech to better understand word form (G)</td>
</tr>
</tbody>
</table>

*Note. The letters correspond to the features of writing (parallel theme) subcodes (see Appendix N): G = grammar; V = Vocabulary; T = Tense; S = Structure; A = Argument; C = Cohesiveness; Su = Support; K = Knowledge.*
Advice Offered

The raters offered four types of advice for the students on how to improve their academic English skills: review and learn, practice, strategies, and tasks. These four types of advice ranged from general with *review and learn* to the most specific with *tasks*. Table 15 outlines the suggestions offered by the raters according to the type of advice. Review and learn, and practice were the only advice types solely focused on language-related features of writing. Strategies and tasks included all three features of writing (i.e., language, organization, and content). Table 16 outlines the frequency of quotations, coded by advice type. Overall, the most common type of advice offered by the raters related to practice. The other three types were offered with similar frequency. Further, as seen in the left-hand column, the number of quotations coded was relatively equal, regardless of the level of essay.

**Review and learn.** The most general type of advice offered by the raters was to review and learn. The advice stemmed from the raters’ comments that the students demonstrated some lack of knowledge and needed to review or learn the material. In other words, this type of advice suggested that the students were missing some knowledge about writing. For the most basic writing, Isobel and Catherine commented on the students’ need to improve their basic sentence structure. For instance, when completing the think aloud protocol for Essay 14, Catherine asked, “Where to start when students have this many mistakes?” (Think Aloud 14) and then reflected “maybe I would start with the basic components of a sentence, making sure that she understands what goes into a sentence, the correct order of a simple sentence compared with a question form” (Think Aloud 14). This comment suggests that it would be best for the student to learn the basic elements of writing before working on the more complex aspects of composing an academic essay. In line with learning the basics of writing an essay, the raters advised the students to learn
the differences between words with similar spellings, such as *affect* versus *effect* or *pet* versus *pests*. Isobel emphasized that the student who wrote Essay 16 needed to understand that the different spellings of similar sounding words could have very different meanings, and that misuse could lead to considerable confusion for the reader.

Raters also identified another skill some students needed to learn: separating essays into meaningful paragraphs. For example, Student 19 had no apparent paragraphing in the essay; Catherine recommended that he needed “to understand paragraphing and where he should begin a paragraph and move on to the next paragraph” (Think Aloud 19).

Table 16

*Overview of Advice Type by Level of Essay*

<table>
<thead>
<tr>
<th>Level of Essays</th>
<th>Review and Learn</th>
<th>Practice</th>
<th>Strategies</th>
<th>Tasks</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>5</td>
<td>14</td>
<td>3</td>
<td>6</td>
<td>28</td>
</tr>
<tr>
<td>Middle</td>
<td>3</td>
<td>10</td>
<td>6</td>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td>Low</td>
<td>4</td>
<td>9</td>
<td>2</td>
<td>8</td>
<td>23</td>
</tr>
<tr>
<td>Total</td>
<td>12</td>
<td>33</td>
<td>11</td>
<td>16</td>
<td>72</td>
</tr>
</tbody>
</table>

**Practice.** The next type of advice type offered to students by the raters was to practice different features of writing. The raters recognized that the students knew the rules and elements of basic writing, but needed to practice the rules in order to refine the skill. All of the suggestions offered by the raters corresponded to language features of writing, but were general enough to be taken up through a variety of different practice activities. For instance, Isobel recommended that Student 17, who was trying to write complex sentences, but obviously struggled, needed “to practice some basic structuring. [He] could do some structure editing exercises to identify main clauses and see how they connect with the subordinate clauses of the sentences” (Think Aloud
This type of advice translates into many different kinds of activities that Student 17 could do, whether completing workbook exercises, examining his own writing for similar types of errors, or working on combining simple clauses. With one high-level student, Catherine advised that he or she practice some tasks to work “on their prepositional usage and the differences from one preposition to another” (Think Aloud 8).

**Strategies.** The third type of advice offered focused on strategies that the students could employ to avoid mistakes. In a sense, the strategies suggested were similar to those that would be offered in a writing course or test preparation class. In contrast to the previous two advice types discussed, strategies included organization and content features of writing.

Catherine offered an organization-related strategy for Student 4. While she commended Student 4 for his clear language, she felt it was a frustrating essay to read since the student took almost the entire writing booklet to describe what he was going to write about, rather than actually writing the body of the essay. According to Catherine, Student 4 needed “to shorten his introduction to get on with the body so that he could . . . spend more time providing details from the readings [to support] what he’s saying” (Think Aloud 4). Most teachers would advise students preparing to write an essay to be concise when writing an introduction and be conscious of how much time they allocated to writing the introduction.

The only strategies offered to low-level students related to spelling. Catherine cautioned Student 19 to double-check the spelling of keywords. The student had spelled “world” incorrectly as “word” throughout the essay. While Catherine acknowledged that the error was relatively minor, it was nevertheless an irritating mistake. Being aware of the keywords is useful for any student, not only when approaching a test task like the CAEL essay, but also when completing a university assignment.
Tasks. The final type of advice provided by the raters centered on the specific activities or tasks that the students should do to aid in their learning of academic English writing. A common task put forth by Isobel was for the students “to create a plan to help them organize their thoughts” (Think Aloud 11). According to Isobel, “the first step would be to decode the topic to really understand the prompt, the second to make a list of information [needed for the essay], and the third would be to make a plan” (Think Aloud 22). Isobel elaborated that doing such tasks would help the students to think about the components of an academic essay to give the writing “some consistency.” Going through the steps of creating a plan targets the organization features of writing, specifically the structure subcode.

For a content-related task, Catherine suggested that Student 18—who struggled with understanding content—not only read more but also practice answering and asking comprehension questions. For improving language, a common suggestion was that the students make charts or lists to help them with verb and noun forms. For instance, Isobel advised that a high-level student, who wrote a coherent essay but seemed to use incorrect verb forms, “make a chart of word forms and the corresponding sentences to see how the words fit into language” (Think Aloud 7). Such tasks target the language weaknesses of confusing different types of word forms and comprehending vocabulary in context. Another language-related task focused on comma usage. Catherine recommended that Student 4 take a long list of parallel words and insert the commas in the appropriate places. According to Catherine, the student “needs to have commas in place which helps your reader out” (Think Aloud 4). The level of detail in the raters’ descriptions when explaining the task easily translated into a task that students could do on their own, hopefully with little instruction.
Discussion

This chapter examined the diagnostic properties of the CAELD essay from the perspectives of two raters. Since there were only two raters and a relatively small number of essays analyzed, the results and conclusions presented in this chapter are provisional and not generalizable beyond the sample collected. However, by focusing on two raters, I was able to provide a detailed account of how a rater identifies diagnostic subcomponents for writing. The findings also provide insight into how raters’ perceptions of the intended purpose of diagnostic assessment and teaching may influence their interpretation of the text and identified subcomponents. There were three facilitating questions and one research question addressed.

**Facilitating Question 1.1:** How did the raters pinpoint the diagnostic information in the students’ writing ability?

The raters focused on the features of the writing that they personally thought improved or diminished the essay. According to Catherine, grammar was the most important skill that students needed to succeed at university and in life. However, Isobel highlighted organization and content as the key features students should improve. The raters’ personal preferences for the features of writing, along with their different reading strategies, might explain the differences exhibited in the think aloud scores. Catherine demonstrated tendencies as a lenient rater, while Isobel appeared to be a severe rater. The think aloud was an artificial activity, used to gain insight into the raters’ perceptions of students’ writing ability and the diagnostic potential of the CAELD essay. As such, the act of completing the think alouds was likely very different from typical rating behaviour. Nonetheless, the task was fruitful in identifying dimensions of students’ writing ability and connecting the information to meaningful advice.
Facilitating Question 1.2: What were the dimensions of students’ abilities as observed by the raters?

There were three distinct dimensions along the continuum of students’ abilities from weaknesses, developing ability, and strengths. Weaknesses are often contrasted to strengths (Jang, 2005). Indeed, major weaknesses were identified that interfered with the raters’ comprehension of the text, while strengths contributed to overall understanding. However, other identified weaknesses were considered minor, as they did not interfere with the raters’ reading. The minor weaknesses did annoy the raters, and the students would need to improve them in order to be successful writers; however, the minor weaknesses should not be the focus of feedback.

The major weaknesses, combined with multiple smaller weaknesses or developing ability, needs to be the focal point for any meaningful advice for the students. The developing ability represents the space between the weaknesses and strengths, in which the students demonstrated some knowledge of the features of writing but were not able to execute their knowledge. Since the students already have some knowledge of the target skill, drawing on this existing knowledge would be an ideal place to work with students to help them improve. Of concern are the limited number of dimensions identified for the low-level essays, which suggests that the CAELD essay may not be the ideal tool to elicit diagnostic information about students’ writing below a threshold of a 30 band score. (The 30 band score represents the cut-off between the low- and mid-level group of essays.)
Facilitating Question 1.3: What advice did the raters offer to the students on how to improve their academic writing skills?

There were four different types of advice, ranging from general to specific, offered to the students by the raters: review and learn, practice, strategies, and tasks. Unlike the dimensions of writing, the raters were able to offer the same frequency of advice for all levels of essays. This similarity in the amount of advice offered suggests that, even though the raters did not identify dimensions in the students’ writing in the same way, they were able to offer some advice on how the students could improve their skills.

These three facilitating questions inform the research question asked in Phase 1:

Research Question 1. What are the features of CAELD scoring that enable it to be used for diagnostic purposes?

I discuss the diagnostic features of the CAELD scoring in relation to grain-size, content, and intended use, three of the defining features of diagnostic assessment described in Chapter 1. First, the raters were able to identify more strengths and weaknesses with a greater degree of detail in the students’ essays above a threshold of a band score of 30. This variation in the raters’ descriptions across the band scores suggests that the higher scored essays allow for more detailed feedback about the students’ abilities. One explanation for this increase in detailed comments is due to the fact that the essays are often longer at a higher band score. As a result, the raters had more writing to comment on. The same degree of variation across the band scores was not observed in the advice offered by the raters.

Second, the majority of the strengths and weaknesses described by the raters corresponded to language and organization categories from the CAELD scoring guide. Catherine and Isobel varied slightly in their approaches to how they determined students’ strengths and
weaknesses and in their preferences for different features of writing, which might have impacted their scoring of the essays. Isobel favoured coherence and content features of the CAELD essays and demonstrated tendencies of a severe rater. Catherine preferred language features of the writing and showed tendencies of a lenient rater.

Third, the raters offered four types of advice about the students’ writing. With the limited degree of specificity about the students’ strengths and weaknesses for the lower scored essays, future research should examine whether or not it would be beneficial to provide all of the students in this lower range the advice for that level. It is likely that most students in the lower range would face similar challenges in their writing.

As with any research, there are limitations that need to be addressed for Phase 1 of the study. First, I would like to acknowledge that the two sources of data informing this phase were indirect measures of the raters’ decision-making processes and conceptualization of the CAELD essays. Unfortunately, there are no direct measures that can access such a complex and cognitive process. By drawing on interviews as well as think alouds, I was able to gain a deeper and richer picture of the phenomenon, than if only drawing on think alouds. Second, the qualitative data were represented both numerically and descriptively. The act of coding qualitative data is grounded in a socially constructed perspective, but the numeric representation of data may be misleading in regards to the precision of the quotation counts. However, I only drew on the frequency counts to look for major trends across the data as a way to complement the descriptive analysis and representation. Third, there were inconsistencies between the original band scores and the think aloud scores. To resolve the consistency concerns, the grouping of the essays into the three levels resulted in only 2 of the 22 essays (or 9% of the essays) having a level difference, rather than the previous 9 of 22 essays (or 40% of the essays). These concerns of score
consistency are discussed in more detail in the data analysis section of the methodology in Chapter 3.

In order to enhance diagnostic assessments, there needs to be more research investigating the nature of students’ strengths and weaknesses beyond the typical dichotomous view. Further, for such assessments to gain traction in the language-learning classroom, the feedback should offer meaningful advice for the students’ next steps for studying.
Chapter 5: Teaching

My teaching philosophy is that I think all assessment, whether it is formative, diagnostic, or summative it should be formative . . . Even if it is the last day of class and I never see them again, I should be looking at what was successful, what wasn’t, and making those changes for later on. It should always inform my teaching. (Diane Interview 1)

Overview

This quotation captures the spirit of diagnostic assessment as a formative tool that informs all aspects of teaching. As such, the purpose of the chapter is to examine the extent to which Diane—the sole participant for this second phase of the study—was able to integrate the Canadian Academic English Language Diagnostic (CAELD) into her classroom instruction. I chose to focus on one teacher’s approach in order to illuminate the complexity involved in the process of weaving a diagnostic assessment—the administration, creation of profiles, and delivery of the feedback—into classroom instruction. In this chapter, I make the distinction between the CAELD as a tool and the diagnostic assessment process. With the CAELD I only refer to the results gained from the students’ performance on the test. With the diagnostic assessment process I refer to Diane’s decisions regarding the implementation the CAELD into her instruction. The learner profiles are diagnostic profiles that were created for each student, using the CAELD feedback for the skills of reading, listening, and writing. (See Appendix E for an example of the profile given to the students, and Appendix D for the ones used by Diane.)

This chapter is organized into three parts. First, I review the research question and facilitating questions and outline the themes and codes informing Phase 2 of the study. Second, I report on the findings as they correspond to the themes and codes. I conclude this chapter with a discussion of the findings in relation to the facilitating questions and research question.
Research Question

The research question addressed in Phase 2 was:

2. What are the teaching conditions that enable the CAELD results to be used for diagnostic purposes?

This research question was further expanded through the following facilitating questions:

2.1. How did the teacher incorporate the diagnostic feedback into her course overall and for individual students?

2.2. How did the teacher interpret the assessment information as diagnostic feedback?

2.3. What aspects of the diagnostic feedback mapped onto the teacher’s overall course planning?

2.4. How did the diagnostic feedback map onto the teacher’s values about teaching, learning, and assessment?

Themes and Codes

There were four main themes for Phase 2 (Table 17). The first theme, background and context, forms the description of Diane in the participant section of Chapter 3 as well as the background and context in this chapter. As the name suggests, the second theme, decision-making process, outlines the choices Diane made for incorporating the CAELD feedback into her classroom instruction. The third theme, using the CAELD, describes how Diane interpreted the CAELD and how it mapped onto her course goals. The fourth theme, spirit of assessment for learning, examines the nuances of how Diane used the CAELD to support her students’ learning and improve her own practice as an instructor.
Table 17

**Phase 2 Themes and Codes with Frequencies**

<table>
<thead>
<tr>
<th>Phase 2 Themes and Codes</th>
<th>Total Frequency (356)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theme 1: Background and context</td>
<td>93</td>
</tr>
<tr>
<td>Code 1: Background</td>
<td>3</td>
</tr>
<tr>
<td>Code 2: Course sections</td>
<td>8</td>
</tr>
<tr>
<td>Code 3: Experience with diagnostic assessment</td>
<td>16</td>
</tr>
<tr>
<td>Code 4: Approach to assessment</td>
<td>24</td>
</tr>
<tr>
<td>Code 5: Approach to teaching and learning</td>
<td>14</td>
</tr>
<tr>
<td>Code 6: Learning expectations</td>
<td>12</td>
</tr>
<tr>
<td>Code 7: EAP course goals</td>
<td>16</td>
</tr>
<tr>
<td>Theme 2: Decision-making process</td>
<td>121</td>
</tr>
<tr>
<td>Code 8: Planning</td>
<td>24</td>
</tr>
<tr>
<td>Code 9: Evaluation</td>
<td>17</td>
</tr>
<tr>
<td>Code 10: Skills and strategies</td>
<td>46</td>
</tr>
<tr>
<td>Code 11: Instructional support</td>
<td>34</td>
</tr>
<tr>
<td>Theme 3: Using the CAELD</td>
<td>29</td>
</tr>
<tr>
<td>Code 12: Diagnostic competence</td>
<td>11</td>
</tr>
<tr>
<td>Code 13: The CAEL as a diagnostic assessment</td>
<td>18</td>
</tr>
<tr>
<td>Theme 4: Spirit of formative assessment</td>
<td>113</td>
</tr>
<tr>
<td>Code 14: Student perceptions</td>
<td>18</td>
</tr>
<tr>
<td>Code 15: Student engagement</td>
<td>26</td>
</tr>
<tr>
<td>Code 16: Self-directed learning</td>
<td>29</td>
</tr>
<tr>
<td>Code 17: Refined teaching practice</td>
<td>11</td>
</tr>
<tr>
<td>Code 18: Change in future teaching</td>
<td>29</td>
</tr>
</tbody>
</table>
Background and Context

The integration of the CAELD into Diane’s classroom instruction was complex. To understand her decision-making process, interpretation of the CAELD, and the nuances of how she used it as a formative tool, it is first important to understand Diane’s approach to teaching, learning, and assessment, as well as the teaching context.

Approach to teaching and learning. The discussions about Diane’s values and her approach to teaching and student learning are intertwined in the next few paragraphs since, in most of the interview data, Diane did not make a formal distinction between the two. Key to Diane’s teaching was her apparent love for learning. Even after six years of teaching, Diane was driven to continually improve her teaching and engage with the overall learning process for her and her students. She mentioned that she liked working within a university environment because of the “opportunity to work with professors [and] benefit from their expertise.” In the first interview, I sensed Diane’s enthusiasm and eagerness to start the new term, because she knew she would have a different set of challenges to overcome; so much so, that she stated, “I love the fact that every time I have a new class it’s a challenge … I feel like I am always learning” (Diane Interview 1). Though she admitted that there were times when she felt overwhelmed with planning and marking, these drawbacks were minimal compared to the benefits that came with teaching and the joy of watching students improve and succeed in their mainstream university classes.

Because of her love for learning, Diane brought energy to her classroom and worked to create a positive learning environment for her students. In the first interview, she described herself as an entertainer, motivator, and facilitator for her students’ learning. To help the students understand the assignments or answer any questions they had, Diane would meet with the
students outside of class time. This desire to engage students was present throughout all the data collected in this phase—interviews, classroom observations, meetings, and documents. For instance, at the beginning of most of the 8:30 a.m. classes, Diane would tell a joke about her love for hockey or some other aspects of her life to break the ice and to help wake up the students.

While Diane was committed to being enthusiastic and motivating in her teaching, she also expected that the students do their homework, attend class, and be active participants in class discussions. Attendance was considered mandatory: Students who were more than 30 minutes late were considered absent, and 2% was deducted from their final mark. Diane summarized her teaching philosophy as well as her high expectations for the students in the course outline, under *instructional approach*:

As your instructor, I believe your success is a shared responsibility. This means your success in [the intermediate-level course] depends on your effort and willingness to commit to the demands and course work, which I’ve established for this level. My responsibility is to develop and support your English Language needs in an academic context with materials and assignments that are interesting, relevant, and challenging. I am committed to helping you succeed, but you must understand that learning is a 2-way street and your successful outcome depends on your hard work and perseverance. (Course Outline, p. 2, bold and italics original)

Though this may seem like a strong approach to teaching, most of the students responded positively by engaging in the class discussions, despite their limited language skills. Diane was aware of the realities of teaching and even suggested, in a meeting, that it was not possible to
reach all of the students. Nonetheless, when some of her students displayed apathy towards the 
learning process, Diane tried continually to engage them.

**Approach to assessment.** At the beginning of the term, Diane was excited to see how the 
term would unfold and looked forward to using diagnostic assessment in her class. She was keen 
to integrate the diagnostic feedback into her instruction, thus having the assessment “benefit the 
students, rather than only placing them into a course” (Diane Interview 1). Informed by her 
previous experience of receiving diagnostic feedback and not using it to help the students, Diane 
was committed to both using the feedback to inform her instruction and sharing it with the 
students during the present study.

Although Diane had six years of experience as a teacher, she was unsure about the 
meanings of, and differences among, assessment concepts. To try to understand the nuances 
Diane pondered, in the first interview: “I think diagnostic is a formative type of assessment, 
right? ’Cause it informs teaching” (Diane Interview 1). In asking this question, it seemed like 
Diane was trying to work through the distinction between summative and formative assessment 
and how diagnostic assessment would map onto them, as well as the possible implications for her 
teaching. Diane’s struggle to articulate the range of assessment possibilities and how she 
embedded them into her teaching was also evident when she described her assessment 
philosophy. Diane’s feelings of being an imposter seemed to stem from her uncertainty about her 
summative assessment practices, as well as from seeing these assessment practices as distinct 
processes from teaching. For her, summative assessment “has to happen. Every term, I try to do 
something differently because I am never satisfied with how I assess … never” (Diane Interview 
1). In addition to vocalizing her struggle to use summative assessment effectively, Diane 
questioned its necessity and its ability to motivate students: “I think all testing should be
formative, but I did say earlier that some students need summative or see it as summative, maybe I don’t know” (Diane Interview 1).

Even though Diane seemed to be frustrated with how to adequately use summative assessment in her teaching and, at times, to question the assessment’s necessity, she was conscious of being fair and accurate in her evaluation of students’ abilities. Diane commented that she would mark most of the assignments blind so she could be as objective as possible. Further, Diane was acutely aware of the high-stakes impact that her marking had on students’ lives: The grades would determine if the students could be admitted into the final EAP course and, ultimately, into the university full-time. To make sure that the grades were reflective of the students’ abilities at the end of term, Diane tried to evaluate their strongest work. This growth approach to assessment seemed to be indicative of Diane’s overall approach to teaching, assessment, and learning: She saw the course as an opportunity to maximize learning and herself as a facilitator. She did not want to penalize the students who had learned the most. For her, the ultimate goal of teaching was to see her students become independent and autonomous learners by the end of her course. In theory, the use of diagnostic assessment to inform her classroom instruction aligned with her value system of classroom instruction.

**Teaching context and course goals.** Diane saw it as her goal to prepare the students for the Canadian university-learning environment, regardless of the course level (beginning, intermediate, or advanced). According to Diane, the students were going to be entering different classroom contexts and needed to experience what their future learning experiences were going to be like: The students should be comfortable working in a lab, in a classroom of 500 students where the teachers use clickers and overheads, or in a seminar class (Diane Interview 1). She
elaborated that “[I am not] only giving them language, but I am also giving them skills and strategies to be successful at university” (Diane Interview 1).

As such, learning strategies were integral to Diane’s approach to teaching EAP. She viewed the strategies as a way to help the students compensate for their limited language skills: “[If I can] give them things they are stronger at, then why not, right? Even if it only boosts their confidence, then it's good.” Strategies, such as skimming and scanning for the main idea in an academic text or tips for organizing a piece of writing, were infused in Diane’s teaching of the courses. Overall, it seemed that Diane saw her role not only as a language instructor, but also as a university preparation coach.

This was a unique classroom setting, in which Diane was able to draw on five teaching assistants. Based on the CAELD feedback, Diane decided to divide the class into four targeted groups, led by teaching assistants: writing, reading, listening, and balanced. Diane had two teaching assistants work with the listening group since it was the largest for both course sections A and B. For section B, Diane created and led a high-level balanced group for students whose proficiency level was considerably higher than other students. To coordinate with the teaching assistants, she had a weekly meeting with them.

The Decision-Making Process

Diane made a number of key decisions about when to administer the CAELD and give back the learner profiles, how and when to include the feedback for her instruction, and how to place students in the targeted groups—all of which are represented by planning. Diane also made a number of decisions on how to integrate the CAELD into her evaluation of the students.

Planning. Diane made a number of planning decisions, such as choosing how to administer the CAELD; she also determined the elements that would be included within the
CAELD, including the self-assessment questionnaire and the specific subcomponents to be included in the learner profiles. Another key decision Diane made was to separate students into groups organized by their weaknesses, as indicated by the CAELD feedback. Next she determined when and how the feedback would be delivered to the students, as well as how the feedback would be interwoven into teaching practice—both for her and for the teaching assistants leading the targeted groups.

There were two stages to Diane’s planning of the course. First, she laid out the course structure a month before it started, and second, she planned the individual lessons on a week-by-week basis. Diane finalized the course syllabus in the second week of December 2009. At that time, she knew she wanted to set up a differentiated style of instruction by creating one group for each of the three skills (reading, listening, writing) measured by the CAELD, plus one balanced group that focused on all three skills. Based on her previous experiences, Diane felt it was important to have these targeted groups scheduled during class time, so that the students would view the process as a valued part of the course and their learning. In the previous term, Diane had also used the CAELD to create similar groups who met outside of class time, and she felt that the students “weren’t motivated to attend or do the tasks” (Meeting 1).

**Evaluation.** Diane used evaluation to convey the importance of the CAELD to the students. There were four key components that made up the course content and evaluation (100% of course weight): two topic-based units (50%), one library project (25%), and one portfolio assignment based on the CAELD (25%). The 25% allotted for the portfolio was the sum of 2% submitting in-class tasks completed for six of the eight sessions (a total of 12%); 5% for the third reflection; 5% for a final task that students decided on after consulting with Diane; and 3% for preparation and presentation. The allocation of 25% of the mark was influenced by how much
class time the targeted groups took up: “I wanted the 25% assigned to [portfolio] because it was such a significant part of the class. We spent 25% of the week was targeted groups” (Diane Interview 2). The portfolio assignment description emphasized that Diane expected the students to take an active role in their learning and, ultimately, become independent learners of academic English. The importance she placed on student awareness is apparent: “Reflections that are worth 7% of their overall mark, which is pretty significant, and in addition they’re doing the reflective type of activity every time they’re in their targeted group, which represents a percentage of their mark” (Diane Interview 3).

Overall, Diane was unhappy with her grading scheme for the portfolio assignment. She seemed to feel that the grades that she had to give to the students did not reflect their abilities because the majority of the marks for the portfolio were based on participation. If the students attended the class and completed the task given to them in class, then they received their 2%. In her own words, “the assessment was too easy … I had given them basically points for showing up to each of the targeted groups, they got points for handing [the portfolio] in a pretty little folder” (Diane Interview 3). If she were to do it again, Diane stated that she would “make the assessment of the portfolio based more on ability with 5 percent based on a test” (Diane Interview 3). Even though Diane wanted to promote the CAELD as important through a percentage of the final portfolio grade, she felt, in the end, that the grade did not reflect the students’ abilities or effort.

**Skills and strategies.** The skills and strategies code emerged from the data as a way to describe the specific diagnostic subcomponents and the corresponding activities that Diane created to help the students improve. Diane’s first decision was to choose which of the diagnostic subcomponents she and the target-group teachers would focus on.
Diane realized that she wasn’t going to be able to focus on the students’ individual weaknesses outside of the targeted groups, other than the high-level balanced group. This was evident when I was helping her create an Excel file to group the information from the students’ learner profiles, one Excel sheet for each section. To understand the class as a whole, the students’ names were arranged in rows, according to their targeted group (reading, listening, writing, and balanced) with their skill scores in the columns. In creating the file, we discussed whether or not the students’ specific weaknesses should also be listed. Diane remarked that, “the reality is there is no way that I can group the students that finitely and teach to those needs, but I could group them according to these four areas (reading, listening, writing, and balanced)” (Meeting 3). Once the learner profiles (teacher and student) were created, the profiles were given to the teaching assistants, so they could focus on the students’ individual needs.

To highlight Diane’s decisions, Table 18 provides an overview of the diagnostic subcomponents used from the CAELD as applied in Diane’s instruction. The first column outlines the subcomponents as measured on the CAELD and learner profile. The last two columns indicate the diagnostic subcomponents that Diane used when teaching the whole class or the high-level, balanced group.

**Reading and Listening.** The low reading and listening scores from the majority of the students in both sections shocked Diane. As a result, she stated that she incorporated more reading and listening activities into her instruction for units one and two (the first 10 weeks of the course). In her instruction, Diane included activities that had the students focus on summarizing the gist of the readings and listenings (see Table 18), or understanding if the questions asked “which, when, or why” (Classroom Observation 5). In the first interview, Diane indicated that she felt one factor contributing to the students’ low reading and listening scores
Reading speed came across as a major concern for Diane. In response, she would include a number of timed reading activities into her instruction. For example, she gave the instructions for the annotated bibliography, part of the library project, as a timed activity. The instructions were broken up into two sections on the front and back of the handout. Word counts of 570 and 355 were marked in the left-hand margin for the sections (Document 5). After three minutes, the students were instructed to mark in the reading where they had stopped. Diane then told the students to record on their tracking sheet approximately how many words they read in the 3 minutes. She asked the students to respond, in the comment column of the tracking sheet, to the following questions: “Was this easy or hard for you? What made it easier or more difficult?” (Classroom Observation 4).

Diane included timed readings as part of her evaluation of the students as well. In the following quotation, Diane described her excitement in including such a task in the unit test: “This was brand new and I like it because I think it’s good, as long as it’s not the only way I evaluate them, it was the timed reading” (Diane Interview 2). Throughout the interview data, reading speed came across as a dominant subcomponent that Diane wanted to address in her teaching.

Diane was also concerned that some students’ passive listening style posed an immense barrier to learning. She commented that the teaching assistants for the listening targeted group really focused on helping the students employ strategies to aid in their comprehension of the material, such as note taking or listening for the gist. Within her own instruction, Diane focused on note-taking strategies.
Table 18

*Diagnostic subcomponents targeted*

<table>
<thead>
<tr>
<th>Skills</th>
<th>On the CAELD and Learner Profile</th>
<th>Diane’s Classroom Instruction</th>
<th>Diane’s Teaching</th>
<th>High-level Targeted Group</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reading</strong></td>
<td>Getting the gist</td>
<td>Reading for speed (I1) and focus (M3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Understanding the main idea</td>
<td>Main idea; vocabulary: specific and keywords; think about the five questions; review terms (CO4)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Locating supporting detail</td>
<td>Strategies: predict; look for cause/effect; paraphrase; definitions; skim/scan (CO Notes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Expanding vocabulary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Listening</strong></td>
<td>Getting the gist</td>
<td>Getting the gist (I2; CO Notes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Understanding the main idea</td>
<td>Notetaking (I2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Catching specific information</td>
<td>Strategies: listen for specific words; vocabulary; listen for details; predict (CO Notes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Using diagrams, labels, and charts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Distinguishing relevant and irrelevant information</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Guessing vocabulary from context</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Making connections to other material</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Paraphrasing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Writing</strong></td>
<td><strong>General</strong>: prompting; length; paraphrasing; quoting; punctuating; spelling; citing;</td>
<td><strong>General</strong>: copying, paraphrasing (M3; CO3)</td>
<td><strong>Grammar</strong>: understanding word form and word category (I2; CO5)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Content</strong>: understanding topic; using support (reading and listening); defining key terms; developing opinions/thesis</td>
<td><strong>Content</strong>: using information from multiple sources; developing a plan; thesis statement (M3; CO3)</td>
<td><strong>Organization</strong>: developing a claim (I3); looking for support (I3); developing an argument (I3)</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Language/Grammar</strong>: expanding vocabulary; understanding word forms, transitions, modality, noun use, articles, and plurals; identifying keywords, and verbs; improving structure, cohesion, and clarity; identifying word categories</td>
<td><strong>Language</strong>: spelling (CO2); avoid personal pronouns (CO2); modals (CO Notes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Organization</strong>: structuring an essay</td>
<td><strong>Organization</strong>: structuring an essay (I2); topic sentence (CO1); transitions (CO2; CO4)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Acronyms represent Interview (I); Meeting (M); Classroom Observation (CO); and corresponding number. For example, M3 indicates Meeting 3.
**Writing.** Prior to the beginning of the course, Diane mentioned that she would often place a greater emphasis on writing because it was going to be expected of the students in their mainstream courses. Diane included a number of strategies or tasks to help the students in their writing, such as planning an essay, developing an argument, and locating support in source material (Table 18). She also mentioned that the teaching assistant for the writing targeted group employed worksheets when focusing on specific grammatical features with the students. In her teaching of the high-level, balanced group, Diane would provide the students with activities to help them work on understanding word form endings, developing a thesis statement, and locating support.

**Vocabulary.** Although vocabulary was not a specific section of the CAELD, it was a subcomponent of each of the three skills. Diane focused a lot of her energies on targeting vocabulary in the readings or listenings that students were given. In the second interview, Diane mentioned that though she often focused solely on vocabulary she also tried to integrate vocabulary with other activities whenever possible. Diane developed tasks to work on the students’ abilities to predict the meanings of words. She included this type of activity in her instruction as well as in her evaluation of the students. In the second interview, Diane mentioned that she felt the predictive tasks “separated the students that actually were using and understanding the vocabulary and being able to predict and guess meaning within context” (Diane Interview 2). Therefore, it seemed that Diane further developed diagnostic tasks that she used to identify the students who understood the vocabulary.

Central to Diane’s process of integrating the diagnostic feedback into her practice was deciding which skills and subcomponents to focus on and how best to address them in the classroom.

**Instructional support.** Diane expressed the need to draw on some form of instructional support, whether in the form of teaching assistants or workbooks, to successfully implement such a large-scale diagnostic assessment into her classroom instruction; in this study, the teaching assistants were key to facilitating the targeted groups. Three main factors led to Diane’s confidence in having teaching assistants lead the targeted groups: the teaching expertise and similar pedagogical beliefs of the
assistants, and the weekly meetings. While the course was unfolding, Diane observed the unexpected benefit of collaborating with the teaching assistants, as well as the unanticipated tension of the extra planning needed to have so many extra teachers in a classroom. Since Diane planned to use diagnostic assessment in her future teaching, she offered insight into how she would incorporate it without the help of teaching assistants.

All of the teaching assistants were finishing or had already obtained their Master’s of Arts in Applied Linguistics, and all had previous experience teaching second language (L2) students. Diane remarked that, if the teaching assistants had not been so qualified or experienced, she would have given them little, if any, class time, and believed she “would not find this high-calibre of teachers again” (Diane Interview 2). Her trust in the teaching assistants also stemmed from her view that they all had similar teaching beliefs and would work towards the same goal. Because of this trust and recognition of the teaching assistants’ experience, Diane was careful to not direct the teaching assistants too much, lest they felt their ideas were devalued.

The weekly meetings created the opportunity for Diane to understand the decisions her teaching assistants made, to ensure that they were in line with what she had planned, and to identify what was not working. For instance, in the middle of the course, the teaching assistants noted that the students were having difficulty completing a tracking sheet and a comment sheet about their own progress. Some of the teaching assistants thought that it would be best if the students did not fill out the comment sheet. However, the comment sheet was important to Diane because it would inform her evaluation of the students and serve as an accountability check for the teaching assistants. Since the students had to articulate which subcomponents they had worked on during the targeted session, the teaching assistant was able to see if he or she had adequately communicated the session goal. Further, Diane indicated that the meetings created an unexpected connection between her and the students by allowing an “opportunity for me to make sure that there are no student issues that need to be resolved” (Diane
Interview 2). The meetings allowed her to maintain control over the targeted groups and to hear any grievances the teaching assistants or students had.

Diane was enthralled with the benefits that the targeted groups brought the students. At the beginning of the course, Diane expressed her excitement about having the teaching assistants in the classroom, stating, “The classroom will be just a buzz of activity. How awesome is that? … I think it's great, so it's a very unique opportunity, if we can make it work” (Meeting 2). An unexpected benefit was the collaboration that occurred between Diane and the teaching assistants. She felt it was the ideal situation because of the number of experts contributing ideas on how to help the students target their learning.

Despite all of the benefits, Diane also noted the disadvantages of having the extra help, most notably, the loss of her teaching time. Midway through the course Diane reflected, “I haven’t had enough time and the targeted teaching time is really sucking a lot out of my curriculum time or scheduled time, which I hadn’t expected or anticipated and that’s been tough” (Diane Interview 2). Another tension was the level of organization that was required from Diane to coordinate her teaching with the teaching assistants: “In terms of designing curriculum to incorporate and use these people, it's tricky, it's a whole other level of administration” (Diane Interview 2). While Diane was excited and enthusiastic to have the extra help, she found the extra planning and organization frustrating, especially when it detracted from her own teaching.

Recognizing that she needed some form of support to effectively use diagnostic assessment, Diane described her visions of using such a tool without the aid of teaching assistants. In the subsequent term, Diane would teach an advanced-level EAP course. She wanted to have the next set of students work on tasks that targeted their diagnostic subcomponents. In fact, Diane regretted not organizing the tasks that the teaching assistants created, so she could adapt the tasks for content and ability. As a long-term solution, Diane proposed the idea of having a workbook full of tasks that targeted the diagnostic subcomponents and “as a teacher you could just almost highlight the table of contents instead of doing a
learner profile” (Diane Interview 3). Even though Diane did have teaching assistants to draw on in order to create tasks, the tasks could be supplemented through workbooks.

Teaching assistants were an integral part of the targeted group approach; Diane’s confidence in the quality of these assistants enabled her to release part of her class to them. Weekly meetings with the teaching assistants were also vital, as they helped Diane maintain both control over the targeted groups and a connection to the students. Diane was motivated to have the teaching assistants in her class because of the benefits it brought her students, but was pleasantly surprised by the level of collaboration that resulted. It is unrealistic to have five teaching assistants in class on a weekly basis; however, already developed tasks corresponding to the diagnostic subcomponents could be used in a single-teacher classroom.

**Using the CAELD**

The theme, using the CAELD, reflects how Diane used the CAELD to create the learner profiles, her opinion on the value of using the CAELD as a diagnostic assessment tool, and how the results mapped onto the intermediate-level EAP course goals.

**Diagnostic competence.** Diane’s diagnostic competence emerged as a central element in implementing the CAELD; she drew on her abilities to design assessments, interpret results, and articulate next steps for course planning to create the profiles and targeted groups, and plan classroom instruction.

At the beginning of the course, Diane expressed confidence in her plan to use the CAELD feedback. She attributed her competence to her prior knowledge of the test as a placement test and her experience using CAELD diagnostic information to inform her classroom instruction. For instance, Diane commented, “I have the experience of being able to make a clearer judgment [in using the results]” (Diane Interview 1). The clearer judgment, it seems, helped her make pedagogic decisions based on the evidence that diagnostic assessment offered. Diane’s ability to use the diagnostic results to
inform her teaching was an essential part of the diagnostic process, but to do so, she needed to first understand what the results meant.

Diane demonstrated diagnostic competence when, while developing the learner profiles, she probed beyond easy conclusions about a student’s ability. At one point, Diane stated that she wanted to understand why a student was initially identified as being weak in reading, given that the results contradicted her initial assessment of the student. After studying the number and types of questions the student answered correctly, Diane determined that the student struggled with reading in a timed setting and paraphrasing material into his or her own words. She observed that a number of other students struggled with these issues, and thus included speed-reading and paraphrasing tasks in her teaching.

To support her conclusions and resulting decisions about student strengths and weaknesses, Diane drew on information outside of the CAELD. For example, when creating a profile for a student who was weak in writing and listening, she remarked, “This writing its not very strong … Um, did he do any planning? [Flips the paper]. There is no planning. These are all teachable skills. And that is what I would really want to focus on” (Meeting 2). The decision-making process that Diane went through before ultimately placing the student in the writing group is apparent. Based on her observations of the student in class, she decided that the student’s listening skills were adequate enough that he or she should focus solely on writing, with a note to focus on organization skills like planning. By focusing on only a few subcomponents, Diane identified not only the next steps for the students in their studying, but for herself in her teaching. She made the decision that the student should not focus on listening skills, even though he or she was also weak in that skill.

This diagnostic competence was apparent during the planning stage. In the initial meetings, Diane outlined the subdiagnostic components that she planned to include in her instruction and was resolute regarding the subdiagnostic skills she wanted to exclude. This decisiveness indicates her understanding of both the subskills measured by the diagnostic test and their appropriateness for her plans and for the level of the course. Diane considered some skills measured by the CAELD, such as
paraphrasing, to be more appropriate for the advanced level EAP course. In choosing to use less diagnostic subcomponents than available, Diane explicitly chose to use only a portion of the CAELD reading and listening sections. Therefore, the students did not need to complete the questions on the CAELD related to subcomponents not focused on in the course.

Diane recognized that every assessment includes measurement error. She stated, “No matter what type of test you use or what type of diagnostic you use, it's not going to be absolute. There's going to be all kinds of variables that will make it ... not valid” (Diane Interview 3). Diane did not allude to measurement error specifically, but the quotation demonstrates that she recognized that any test, including the CAELD, is only an estimation of student ability with a number of possible confounding variables. Given the potential limitations of assessments, Diane acknowledged the possible need for other sources of information, in addition to the CAEL, when creating the profiles that would inform her pedagogic decisions.

Diane displayed considerable diagnostic competence and a sophisticated understanding of the different diagnostic subcomponents measured by the CAELD, as well as an awareness of the possible, confounding variables in the assessment process. Further, this competence required a concrete understanding of her needs and goals as a teacher, and of the needs of EAP students and future, full-time university students. The diagnostic competence considered in this study is, therefore, slightly different than the concept as outlined by Edelenbos and Kubanek-German (2004). In their study, Edelenbos and Kubanek-German described diagnostic competence as the teacher’s skills and ability “to use and interpret data from various types of language sources” (p. 260). For the purposes of this study, diagnostic competence refers to the teacher’s ability to understand and synthesize the information identified from one main source, a diagnostic assessment, while also factoring in other variables.

By drawing on multiple forms of evidence in addition to the initial CAELD conclusions, the validity of Diane’s decisions and her judgments of the students’ strengths and weaknesses were most likely enhanced. In addition to enhancing the validity of the learner profiles, Diane formed a deeper
understanding of the students’ needs as a result of being involved in the process of creating the learner profiles.

**The CAEL as a diagnostic assessment.** Diane had mixed perceptions about using the CAELD. Her initial reaction upon giving and scoring the CAELD was that it took too long to administer and to create profiles. She felt that the writing was at an appropriate level, but the reading and listening were overly difficult for intermediate-level EAP students. Despite these drawbacks, a benefit for her was that the CAELD was a validated test compared to one that she would develop by herself.

The CAELD was administered on the second day of the course in each of the two sections. Because the test is three hours in length, it took one full class to administer. Some of the students did finish early, but most took the full three hours to complete the test (Classroom Observation 1). For the students who were absent from the scheduled test day, a make-up session was offered. According to Diane, the three CAELD sections did not need to be as long because only “a snapshot” (Diane Interview 1) was necessary for identifying student strengths and weaknesses. Diane saw that the students were tired from completing such a long task, especially when they got to the last section on writing. She remarked that, as a result, the writing section did not draw out some of the students’ true abilities.

Diane was “really unhappy about the [learner] profile” (Diane Interview 1) because in her opinion they took too long—approximately two weeks—to create. The long timeframe was caused, in part, by the make-up session for students who missed the original administration. That Diane spent so much time creating something that she did not use, as the profiles were ultimately used by the teaching assistants, probably contributed to her frustration with the time-consuming process of creating the learner profiles. Plus the creation of the profiles was inefficient. One step in the process required transferring information about the students’ CAELD scores to the learner profile. Having the raters record such information on the profiles would have eliminated this step. That being said, Diane’s proximity to the results while creating the profiles might have afforded her a deeper understanding of the
students’ abilities. Overall, the length of the CAELD administration and time spent creating the profiles seemed to be the largest sources of frustration for Diane.

Diane was critical of the ability level of the CAELD for her students. Overall, she thought the writing section was a suitable level for the students in the course. Initially, she distrusted the reading and listening results because so many of the students performed markedly lower than Diane anticipated. Upon looking through the CAELD reading section, Diane stated that, “I would expect them to be able to read [the level of reading we gave them]” (Meeting 2). In the first interview, Diane repeatedly stated that she would consider changing the length of the reading, but questioned why the students did not finish: “What is it about the text, why are they not getting through? Is it too complex or is it because it’s actually too long?” (Diane Interview 1). In the second interview, Diane stated that the reading was just too difficult for the low-level students: “They didn’t get any of it.” She expanded, saying that it was most likely a misplacement issue, and not reflective of the difficulty level in the CAELD reading. Nevertheless, Diane was disappointed that the CAELD reading and listening sections were not more accommodating to the wide-ranging abilities of her students.

Frustrated that the CAELD listening was too difficult for the students, Diane chose to create her own listening diagnostic the following term. However, she noted that her self-developed listening diagnostic was too easy and observed that “the students [got] all of the questions right. It’s not doing anything for me either. So finding this balance is difficult, and you know it takes time” (Diane Interview 3). In developing her own diagnostic test, Diane realized the challenge in identifying skill subcomponents: “The one thing I’m missing right now is legitimate test specifications … I really have had to work hard at trying to establish questions that are getting at some of those subcomponents” (Diane Interview 3). The experience of developing her own diagnostic allowed Diane to see the challenge of creating a test at an appropriate level as well as the difficulty in identifying subcomponents such as those encompassed by the CAELD. With her own diagnostic, Diane tried to identify test-specifications and diagnostic subcomponents retrospectively.
Diane was “stuck on whether or not she would change the CAELD” (Diane Interview 2). She recognized the benefits that a validated test brought her. The content was similar to what she would use in her EAP course, but Diane nonetheless appeared dissatisfied with the difficulty level.

The Spirit of Formative Assessment

The spirit in which the CAELD was used is perhaps the most vital element to integrating it as a formative assessment into classroom instruction. Diane was committed to making sure that the CAELD was beneficial for her students. Based on her previous experience using diagnostic assessments, Diane expressed frustration with herself for not including the students more in the process, even by doing something as simple as giving them the results. This regret led her to commit to focus on how the students used the results, and influenced the degree to which she wanted them to be active participants represented by student perceptions and student engagement. Her desired goal for the course and for the use of the CAELD was for her students to become more independent and self-directed learners. Diane also wanted to learn from the experience, in order to refine her practice.

Student perceptions. Diane continually sought to gain a sense of the students’ perceptions on their learning and her teaching during the course. She wanted to know if the students were content with the decisions resulting from the use of the CAELD and with being taught by someone else in the targeted groups.

On the day that the students received the CAELD feedback, Diane walked around the room to see if the students were content with the target group into which they were placed (listening, reading, writing, and balanced) and whether or not they thought the decision was appropriate. Throughout this process, Diane would sometimes have to negotiate with the student about the placement in a particular target group. This was the case with a student who was placed into a listening group, but argued that he was stronger in listening and should be placed in one of the other groups:

Diane: What's interesting for me is that you, in your self-assessment in your questionnaire you identify that you were slightly weaker in listening.
Student: I think listening is little bit higher than the reading and writing. I have to say for me it's my weakness.

Diane: Okay, but in the self-questionnaire you said it was lower than reading and writing.

Student: Oh really?

Student: Yeah, because sometimes a lot of people say I don't understand them because of talking too fast.

Diane: Okay, so do you think if I put you in this group do you think it's gonna help you?

Student: I think it's gonna help me. Because this class is not [only on listening].

(Classroom Observation 2)

Even though the student originally thought he should be placed into a different group, Diane negotiated with the student to stay in the listening group based on his self-assessment questionnaire responses. In this situation, the self-assessment was a communication tool and an additional source of evidence to support the placement decision. Diane used it not only to help herself understand how the students perceived their language ability, but also to help students see value in the decision.

For Diane, making sure that the students were satisfied with the decisions from the CAELD was essential. In a meeting following the class, Diane commented that most of the students trusted her placement decisions for the targeted groups. For those who did not, she offered to look at the profiles again, along with other information, to see if the students should be moved to another group: “I will also go back and look at some of the data that [the administrative secretary] gave me to see the performance in other tests and stuff like that. And that’s fair” (Meeting 4). By checking the congruence between the students’ perceptions of their abilities and their results, Diane sought to confirm the decisions for herself and to ensure that the students had confidence in the results and in the targeted group placement decisions. Diane also tried to gain a sense of the students’ perceptions of the diagnostic assessment process to make sure that they were content with receiving targeted group instruction from someone else. She was aware that the students had paid considerable amounts of money to take the course with
her, so she wanted to ensure that the students did not feel “ripped off” (Diane Interview 2). Ultimately, for her, the students were the customers and she wanted to ensure that they were satisfied.

Throughout the course, Diane informally gathered student perceptions. She did this to see if students’ self-assessments aligned with the CAELD feedback and if they were happy with their target group placement and with being taught by another teacher.

**Student engagement.** Diane surveyed the students’ reception of and progress with the diagnostic assessment by monitoring their engagement with the tasks and activities. This idea of student engagement is congruent with Diane’s approach to teaching, as described at the beginning of the chapter. She expected the students to attend class, actively listen, and contribute to class discussions. Diane extended these same expectations to the students’ participation in the diagnostic assessment process.

Diane considered the students’ engagement with tasks to be a sign of their comprehension of the material, and therefore, of the success of the overall process. For example, Diane commented that she was impressed with the students’ engagement with the reflections “because it was a hard task getting them to look at something fairly abstract” (Meeting 3). Similarly, in describing how the students should use the CAELD feedback, Diane stated that she was “worried that [she] would lose students, that they’d start chatting and be off topic, but I thought [with] both classes that didn’t happen. I really think for the most part they were into it” (Meeting 3). Diane believed that the students demonstrated their comprehension of an activity or task by working through it or actively listening to instructions.

Conversely, Diane considered limited participation to be a possible sign of student disinterest. As stated previously, Diane mentioned that she was not willing to continually try to connect with students who were not motivated to work hard. Diane expected the students to take the CAELD and the entire process seriously: “I can diagnose what is lacking in terms of ability then I can work on teaching that, if it’s just a matter of effort well then too bad, so sad” (Diane Interview 3). This quotation seems to reflect
Diane’s approach that she was willing to help the students as much as possible, but only if they were willing to work hard.

To promote student engagement, Diane consciously paired students who would not normally work together. Pairing students with different personalities encouraged the students who did not normally contribute to class discussion to do so in their targeted groups. She mentioned that, though the students might not have been initially fond of the situation, it was beneficial in the end. To encourage student participation, Diane also used grades as a form of motivation to encourage the students to do the tasks or take the process seriously.

Diane believed student engagement provided evidence with which to determine the success of the diagnostic assessment process, specifically as related to individual activities or the targeted groups. It also served to help distinguish between the students who took the CAELD seriously and those who did not and were not motivated.

**Self-directed learning.** One of Diane’s central goals for the course was for the students to become independent and successful learners in a university context. Diane created opportunities for the students to discover the connections between the CAELD feedback and their own abilities, to target one or two subskills (i.e., reading, listening) by focusing their energies on only a few diagnostic subcomponents, and then to articulate a plan for improving the targeted subcomponents. To help the students become successful, self-directed learners, Diane provided support in the form of checklists and extra practice materials online.

As CAELD feedback was unique to each student, Diane emphasized that using the profiles should be, as well. Immediately before the students received the CAELD feedback, Diane instructed that the feedback was “individual. Don’t be sharing with everyone. Like what did you get … it does not matter. It matters what you got” (Classroom Observation 2). Diane wanted to dissuade the students from competing with each other, and to focus only on their results. Diane also expressed that some students were likely to be more successful at self-direction than others. While describing the high-level targeted
group and why she felt they would be more competent with independent tasks, Diane appraised the students as “highly motivated, … bright, and resourceful. So, give them a task to do, tell them where they can get support and make use of all the services the university has” (Meeting 3). Diane’s view was that successful independent learners were motivated and utilized the resources and support offered to them.

For Diane, preparing the students to be successful self-directed learners included encouraging them to become more aware of their language abilities, strengths and weaknesses, and of the next steps they needed to take to improve. A beginning step in this process was for the students to “set goals” (Diane Interview 3) and narrow their energies on their learning needs, so the target becomes a “shiny object they can focus on” (Diane Interview 3) against the backdrop of all the possible diagnostic subcomponents they could focus on. Thus positive reinforcement combined with successful outcomes was “the carrot” (Diane Interview 3) that the students needed to continue as active learners: “They have been able to make really concrete connections between this is what you did right and this is the outcome” (Diane Interview 3). Diane’s emphasis on the students making connections is apparent in the following excerpt of her instructions to the class, in which she modelled how the students could think about using their strengths to improve their weaknesses:

Diane: How would you use your strengths to improve your weaknesses? So, if reading is my strongest skill then how can I use my reading strengths to improve my writing? No kidding.

Students: [laughing]

Diane: How? By paying attention to certain vocabulary or grammatical structures.

(Classroom Observation 2)

Diane felt that having the students notice and become aware of how they were improving was essential to their success. A key part of this awareness was recognizing that strengths could be a vital asset with which to improve their weaknesses.
Part of this awareness included the need to transfer the targeted diagnostic subcomponent to other situations. In the following quotation, Diane described a student from the high-level targeted group, who was focusing on endings and word categories, and still made a number of mistakes in his or her oral presentation:

You know how I’ve been working on word form and word category and stuff a lot. One of the students is still making a lot of mistakes in the oral presentation. So, I’m actually going to mention that to that person tomorrow because I will have handed back their oral presentation with feedback and I want them to sort of reflect on that like why didn’t they use any of the strategies … why didn’t they transfer that into preparing for their oral presentation? (Diane Interview 2)

In the quotation, it is clear that Diane expected the students to make the connections between practicing the diagnostic subcomponents, in class and through tasks, and transferring this new knowledge and ability to other situations.

Diane was dissatisfied with the limited focus on student strengths in the diagnostic profiles. At the beginning of the term, when talking about the type of information that she wanted to pass onto the teaching assistants, Diane realized that the CAELD did not put “enough emphasis on their strengths as a result of this process” (Meeting 2). She related the need to focus on strengths to how her mother as “an elementary school teacher … and she was always marking the students [work with] how many they got right, you would never say you got three wrong” (Diane Interview 1). In focusing primarily on the students’ weaknesses, the CAELD neglected to incorporate a key component of diagnostic assessment. According to Diane, it was essential to use students’ “strengths to build [on the students’] weakness, which should be the purpose of the diagnostic, [in fact] I think is one of the definitions” (Diane Interview 1).

The use of the CAELD feedback with an assessment as learning pedagogical approach with the students was at the core of the diagnostic assessment process. The data illustrated that it was important
to Diane that the results and, perhaps most importantly, the entire process was meaningful to the students. Diane considered the students’ perceptions to see how they interpreted the results and her instruction. In addition, Diane wanted to observe the degree to which the students engaged in the process; further, she wanted to know whether any lack of engagement was a result of students not understanding or simply not caring. Self-directed learning was one of Diane’s goals for the diagnostic assessment process: She wanted the students to become independent and successful full-time university students, who would be able to identify and draw on university support services to help them improve. Through her instruction, Diane encouraged the students to make concrete connections between their identified needs and progress in their learning.

**Refined teaching practice.** Using the CAELD as a source of diagnostic feedback on students’ abilities resulted, for Diane, in a “refined” (Diane Interview 1) teaching practice and a heightened awareness of her students’ needs. Diane described making clearer judgments about the students’ abilities and deciding on teaching activities, based on the well-defined criteria that using a diagnostic assessment gave her. Further, the diagnostic assessment process allowed Diane to get “to know the students a bit better and [understand] what they really need[ed] because the population has changed, and it continues to change” (Diane Interview 3). The CAELD feedback reminded Diane to stay on track and focus on the aim of the course, instead of basing her instruction only on intuition and previous experiences, lest “you lose sight of what is important” (Diane Interview 3). In light of this directed teaching focus, Diane did comment that the CAELD information might have caused her “to pigeonhole some students” (Diane Interview 3), but felt that this early labelling prompted her to make sure the students received more support than she would have otherwise given them.

Diane’s apparent heightened awareness stemmed from her access to detailed and well-defined criteria. She used the CAELD feedback to set specific objectives, direct her teaching, and give her confidence in her final evaluations of the students. The criteria informed the curriculum and Diane’s teaching by creating a focus: “So I am not just teaching generally to reading. I’m teaching to reading
specs” (Diane Interview 1). Having a curriculum tailored to [or informed by] the needs of the students influenced her feedback and evaluations. According to Diane, she was able to “articulate [her] expectations more clearly to the students and maybe feel a bit more sure that a student [was] prepared to more forward” (Diane Interview 3). The use of the CAELD led to an enhanced awareness in Diane’s teaching. The results persuaded her to continually consider the goal of the course when designing assignments and tasks. Further, the CAELD outlined criteria that Diane could use to define her underlying course curriculum.

**Influence on future teaching.** The use of the CAELD seemed to have an effect on Diane’s ongoing teaching practice. This impact was apparent during Diane’s third interview, where she described how she planned the course she taught immediately after the one considered in this study. Diane commented that, in her future teaching, she would have a clearer understanding of what to focus on. In addition, she alluded to her plan to keep using a form of diagnostic assessment in her courses, although it would be modified from what was used in the study, and most likely not based on the CAELD. Though she valued having the students reflect on their abilities and on the next steps that they planned to take, Diane intended to incorporate less reflections into future courses.

**Discussion**

Administering a diagnostic assessment, developing learner profiles, and then using the results to inform instruction is a complex, exceedingly idiosyncratic process. Underlying this endeavour are the values inherent in one’s approach to learning, teaching, and assessment, and the relationships among them. This chapter highlights Diane’s approach to integrate the CAELD into her classroom instruction. The key findings are discussed in relation to the four facilitating questions guiding the chapter. These questions, listed below, explore how Diane incorporated the diagnostic feedback into her classroom instruction and interpreted the CAELD as diagnostic information, as well as how the feedback mapped on to the course curriculum and Diane’s values of classroom instruction.
Facilitating Question 2.1: How did the teacher incorporate the diagnostic feedback into her course instruction overall and for individual students?

For the CAELD to be a meaningful and useful classroom pedagogic tool that impacted student learning, Diane felt that it had to do more than diagnose students’ strengths and weaknesses: Diane wanted to make sure she used the results to benefit the students. To do this, Diane drew on the CAELD as a formative assessment tool. McMillan (2010) described the use of formative assessment as having three levels (e.g., high, moderate, and low) with 10 defining characteristics: evidence of student learning, structure, participants involved, feedback, timing, instructional adjustments, choice of task, teacher-student interaction, role of student self-evaluation, motivation, and attributes for success. Drawing on McMillan’s (2010) three levels of formative assessment seen in Table 3, Chapter 2, I now discuss how Diane’s use of the CAELD demonstrated that it had characteristics of both a low-level and a moderate-level formative assessment.

Low-level formative assessment. The CAELD is a standardized test that is typically administered in large-scale settings. By default, the standardized characteristics of the test gives it features of a low-level formative assessment, namely, evidence of student learning, structure, and limited choice of task. The CAELD is a criterion-referenced assessment that is administered under timed conditions; students are not allowed to discuss their answers with one another. These features support the reliability of the inferences made about the students’ abilities. However, because of the standardized administration conditions of the CAELD, it is administered in a high-stakes, test atmosphere; this testing environment detracts from the formative spirit of a learning culture (Shepard, 2000) in which assessments should be given in low-stakes settings. To maintain the integrity of inferences made about the students’ performance, the standardized administration is unavoidable. As a standardized assessment, the CAELD had to have a formal structure within the course. The CAELD was chosen because of its close history with the program. Nevertheless, the test was not developed for Diane’s
course, nor did she have a role in choosing it as an assessment tool; thus it had qualities of a low-level formative assessment.

Based on the instructional adjustments and feedback characteristics (McMillan, 2010), the pedagogic decisions that Diane made using the CAELD feedback would also be classified as low-level formative assessment. The decisions made were structured and well planned: Diane planned the portfolio assignment and the use of targeted groups even before the students took the CAELD. When Diane did organize the targeted groups, she looked across the two sections of CAELD scores and considered the students’ individual personalities and abilities. By taking the time to ponder the students’ needs, Diane created the unique, high-level group. In doing so, she ensured that both the students in the regular targeted groups and those in the unique group were given the opportunity to have their needs met. Being placed in a targeted group had moderate stakes for the students. For example, highly motivated students might have preferred to be placed in a different group and focus on a different skill. The CAELD feedback was provided to the students two weeks after they had completed the assessment, thus making it a low-level formative assessment. Diane did make changes to her teaching about halfway through creating the learner profiles.

**Moderate-level formative assessment.** According to McMillan’s (2010) table, there are two characteristics that place the CAELD as a moderate-level formative assessment: role of student self-evaluation and attributes for success. To create the CAELD learner profiles, the student self-evaluations were consulted and incorporated into the feedback, or in McMillan’s terms, they were “tangential” (p. 43). Teacher-student interactions played an important role in Diane’s communication of the CAELD feedback to the students. Through these interactions, Diane was able to gather the students’ perceptions about both the CAELD feedback and their placement in the targeted groups. Further, she was able to promote self-direction by encouraging the students to think about their own learning. In the end, the teacher-student interactions involved some negotiation, but Diane’s was the dominant voice. Though Diane was interested in the students’ perceptions, she generally persuaded the students to accept the
CAELD feedback and placement decisions. This teacher-directed consultation of results limited the extent the students’ could have self-direction and regulation.

Student motivation is an important feature of formative assessment, especially in the way Diane used the CAELD. Diane incorporated features into her practice to help motivate students. First, she gave the CAELD a prominent role in her instruction by administering it during class time: She wanted the students to take it seriously. Second, the portfolio assignment accounted for one-quarter of the overall grade. Though the only mark attached to the CAELD was a participation mark for attending class, there were significant marks attached to the resulting assignment. Third, in her individual teaching of students, Diane encouraged the students to use a tracking sheet to monitor their progress and reflect on how they could improve.

**Facilitating Question 2.2:** How did the teacher interpret the assessment information as diagnostic feedback?

Since Diane created the learner profiles, she was deeply invested in considering how the CAELD performance was diagnostic for her students and for herself. In both cases, she looked to the self-assessment to see if the students were accurate in their own assessment of their abilities. Diane also reflected on the results for her own teaching practice. To incorporate the CAELD feedback into her instruction, Diane interpreted the students’ CAELD performance first by skill score, course section, and student, and then by diagnostic subcomponent.

Based on the number of students with low reading and listening scores on the CAELD, Diane chose to add a number of reading and listening activities to her instruction, which she had not originally planned to include. Though reading speed was not a diagnostic subcomponent in the CAELD, Diane interpreted some students’ low performances on their reading, combined with their inability to complete all of the questions, as an indication of weakness in reading speed. This interpretation led her to incorporate a number of tasks targeting the skill into her instruction.
Underlying Diane’s interpretation of the CAELD was her diagnostic competence. This competence seemed to be rooted in Diane’s previous experiences and confidence in teaching. As a result of the diagnostic competence, Diane was decisive about how to use the CAELD feedback.

**Facilitating Question 2.3:** What aspects of the diagnostic feedback mapped onto the teacher’s overall course planning?

Diane noted that the tasks included in the CAELD were similar to tasks that she used in her course. The CAELD writing section was most useful for Diane’s instruction because of the type of information it elicited as an open-ended task. However, she did not feel that the CAELD reading and listening sections were at an appropriate level of difficulty for the students. She felt that, if it had been easier, more of the students would have been able to answer the questions. Diane was able to incorporate the CAELD feedback across the course assignments: the two units, the library project, and the portfolio assignment. Using the CAELD feedback, Diane encouraged student independence—one of her main course goals—through reflective activities, self-directed tasks, and instruction.

Although Diane recognized the importance of oral skills in an EAP classroom, she did not want to include oral skills into the diagnostic assessment process. Beyond improving academic language skills, the course was designed to help the students acculturate to the Canadian university environment. Part of this acculturation process involves interacting with peers. Interactional competence is a feature not measured by the CAELD. However, Diane’s decision to use the results to create subgroups forced the students to work with peers with whom they would not otherwise have had the chance to work.

**Facilitating Question 2.4:** How did the diagnostic feedback map onto the teacher’s values about teaching, learning, and assessment?

The opening quotation for this chapter is from my first interview with Diane. She described her philosophy of assessment and the importance of using it as a continual source of evidence. This desire to learn was apparent in the section describing Diane’s approach to teaching, learning, and assessment. In a
similar fashion, Diane used the CAELD as a source of evidence for her own practice, and for improving her students’ academic language development.

Using the CAELD was also a learning opportunity for Diane—one that helped refine her teaching practice and influence her future teaching. She used the diagnostic descriptors to shape her course curriculum, feedback, and evaluations of students. While Diane was aware of these advantages, she was also deeply critical of the CAELD. She felt it was too long; she believed a shorter, less gruelling assessment would have garnered a more accurate estimation of the students’ writing abilities. Further, she found creating both learner and teacher profiles to be a redundant and time-consuming job.

There were three key features of her pedagogic decisions based on the CAELD that Diane valued: (a) the independence fostered in the students, (b) the combination of different peer groups, and (c) the perceived high-level of student engagement. In contrast, there were three key aspects of the decisions made that did not align with Diane’s values: (a) her access to the students, (b) her worry that she was not their primary teacher, and (c) the limited focus on student strengths. Diane noted the more motivated students were also the ones who benefited the most, which was not necessarily counter to Diane’s value system of student learning.

These four facilitating research questions inform the research question asked in Phase 2:

**Research Question 2**: What are the teaching conditions that enable the CAELD results to be used for diagnostic purposes?

The teacher and the supportive teaching environment were the critical factors that enabled the use of the CAELD. Diane’s diagnostic competence, involvement in the scoring process, and commitment to fostering a positive learning environment grounded in principles of formative assessment were essential to the process. The teacher’s integral part in the creation of the profiles allowed her to see the connection between the students’ performances and their identified skills. When possible, she reviewed the results to ensure the profiles were appropriate and matched her own assessment of the
students’ abilities. As such, the decisions made by another teacher could be very different, even within the same program and diagnostic information.

The instructionally rich context enabled differentiated instruction that would not have been possible without the teaching assistants. Support provided to the teacher proved essential to the use of the CAELD for diagnostic purposes. The development of targeted teaching activities, small group teaching by the teaching assistants, and administrative support provided by myself, the researcher, allowed Diane to maximize the potential of the CAELD for diagnostic purposes.
Chapter 6: Learning

*I think the feedback . . . is very useful because I know the score every time I do the listening, and it’s not very good. I can’t raise the score and that means I don’t know which aspect I need to practice in listening . . . Then if you give me the comments . . . I know oh this is my weakness, so I need to focus the main aspect to stress it. To do more practice for listening skills. This is very, very important, importanter than school I think.* (Douglas Interview 1)

*It’s good to get the diagnostic feedback and know the levels of the students, and the skills, and what are the weaknesses. But, the issue is how to deal with this. How to solve the problem? Should the teacher give the students practice writing a lot, or just [give] it to you, but you have to solve it.* (Dennis Interview 1)

**Overview**

The above two quotations, from two of the students interviewed, illustrate how students’ use of diagnostic feedback varied in relation to their learning goals and unique classroom experiences; for instance, while Douglas believed the feedback could help improve his personal studying habits, Dennis questioned the roles and responsibilities of the teacher and student. This chapter examines the individuality and the similarities of the event. To achieve this goal, this chapter investigates how nine students interpreted and made use of the CAELD feedback, and considers their advice for improving both the feedback and the targeted instruction as aids to studying academic English. The 47 students’ reflections—given at the beginning, middle, and end of the course—facilitate a comprehensive understanding of the phenomenon. This chapter reviews the research question and facilitating questions that guided this phase of the study, reports on the themes and codes that emerged from the data, and discusses the findings.
Research Question

The research question addressed in Phase 3 was:

3. What are the learning conditions and students’ perceptions that enable the students to use the CAELD and its results to support their learning?

This research question was further expanded through the following facilitating questions:

3.1. How did the students use the diagnostic feedback to benefit their learning?
3.2. How did the students interpret the CAELD feedback?
3.3. How did the diagnostic feedback map onto the students’ learning goals?
3.4. How did the diagnostic feedback map onto the students’ values about learning and assessment?
3.5. What was the students’ test taking experience of the CAELD?

Table 19

Phase 3 Themes and Codes with Frequencies

<table>
<thead>
<tr>
<th>Phase Two Themes and Codes</th>
<th>Frequency (443)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Theme 1: Experiencing the CAELD</strong></td>
<td>56</td>
</tr>
<tr>
<td>Code 1: Administration</td>
<td>21</td>
</tr>
<tr>
<td>Code 2: Perception of the CAELD</td>
<td>35</td>
</tr>
<tr>
<td><strong>Theme 2: Experiencing Diagnostic Teaching</strong></td>
<td>89</td>
</tr>
<tr>
<td>Code 3: The learner profile</td>
<td>30</td>
</tr>
<tr>
<td>Code 4: Perceptions of teaching</td>
<td>44</td>
</tr>
<tr>
<td>Code 5: The reflections</td>
<td>15</td>
</tr>
<tr>
<td><strong>Theme 3: Targeted Learning</strong></td>
<td>298</td>
</tr>
<tr>
<td>Code 6: Reactions to and use of feedback</td>
<td>54</td>
</tr>
<tr>
<td>Code 7: Perceived strengths and weaknesses</td>
<td>67</td>
</tr>
<tr>
<td>Code 8: Evolving conceptions of diagnostic assessment</td>
<td>47</td>
</tr>
<tr>
<td>Code 9: Learning goals</td>
<td>43</td>
</tr>
<tr>
<td>Code 10: Approaches taken to target learning</td>
<td>87</td>
</tr>
</tbody>
</table>
Themes and Codes

Twelve codes emerged from the data, forming the three themes of *Experiencing the CAELD*, *Experiencing Diagnostic Teaching,* and *Targeted Learning*. Table 19 outlines the overall frequency counts of the three themes and the individual counts for each code.

**Background and Context: Students Interviewed**

Of the 47 students who participated in Phase 3, 9 were subsequently interviewed. The nine students interviewed had similarities and differences in their English language learning experiences. Below, I describe each student’s length of time in Canada, age, home country, first language, CAELD scores, target group, and unique English learning experiences.

*Elaine* had lived in Canada for two and a half years prior to the study, and was 20 years old. She had spent two years studying at a Canadian high school, and had taken the beginner EAP course the previous semester. Elaine was from Bangladesh and spoke Bengali as a first language. Elaine’s CAELD scores were 50, 71, and 76 for writing, reading, and listening, respectively. Elaine was placed into a balanced group.

*Laura* was a mature Iranian student, aged 52, and spoke both Kurdish and Persian. She reported practicing her English outside of the program by completing practice exercises on her own and speaking English with her husband, who was also Iranian and was completing his PhD at the same university. Laura’s CAELD scores were 50, 39, and 43.75 for writing, reading, and listening, respectively. Laura was placed into a reading group.

*Douglas* had lived in Canada for a year and a half at the time of the study, and was 22 years old. The intermediate EAP course was the first for him in the program, but he had taken other language support courses in Canada. Douglas’ CAELD scores were 60, 66, and 43.75 for writing, reading, and listening, respectively. Douglas was placed into a listening group.

*Neil* had only been in Canada for 3 months when the study began and he was 23 years old. Neil was determined to obtain the CAEL band scores in order to be admitted to the Master’s of Economics
program at the university. Neil took the CAEL again a month before the end of the study. During the second interview, Neil exclaimed that his latest CAEL scores satisfied the language requirements for the master’s program. In the first interview, Neil attributed his strong English skills to speaking English with his stepfather, who was Canadian and spoke English as a first language. Neil’s CAELD scores were 50+, 72, and 59 for writing, reading, and listening, respectively. Neil was placed into a balanced group.

Dennis had been in Canada for a year and was 31 years old. He was from Saudi Arabia and spoke Arabic as a first language. Dennis had taken the beginner level EAP course the previous semester. Dennis’ CAELD scores were 40 and 45 for writing and reading, respectively. There was no score for listening, as Dennis was unable to complete that section of the test. He was placed into a writing group.

David had been in Canada for a year and was 19 years old. He was from Saudi Arabia and spoke Arabic as a first language. David took the beginner EAP course the previous semester. He reported having spent time in other English-speaking countries through travelling and summer school. David’s CAELD scores were 30, 29, and 37.5, for writing, reading, and listening, respectively. He was placed into a writing group.

Stephen had been in Canada for two years and was 21 years old. He was from Saudi Arabia and spoke Arabic as a first language. He had taken courses similar to the intermediate-level EAP course at another university in Canada. Stephen’s CAELD scores were 30, 30, and 36 for writing, reading, and listening, respectively. He was placed into a balanced group.

Don had been in Canada for four months and was 17. Don was originally from Egypt and spoke Arabic as a first language. His father was a diplomat and, as a result, Don had lived in four other countries (Kenya, Philippines, Spain, and the United Arab Emirates) before moving to Canada. He started to learn English when attending school in Kenya and continued in the Philippines. Don had taken the beginner level EAP course the previous semester. His CAELD feedback scores were 60-, 76, and 61.25 for writing, reading, and listening, respectively. He was placed into a balanced group.
Kevin had been in Canada for more than a year and was 21 years old. He was from Saudi Arabia and spoke Arabic as a first language. He had taken language support courses at a community college in Canada. Kevin’s CAELD scores were 50, 39, and 54 for writing, reading, and listening, respectively. He was placed into a reading group.

**Experiencing the CAELD**

The students’ experiences were grounded in their actual encounters with taking the CAELD for the purpose of identifying their strengths and weaknesses for the intermediate-level EAP course, which is described in detail in the *Administration* code. This event was further shaped by how the students perceived the CAEL or CAELD based on prior knowledge.

**Perceptions of the CAEL/D.** The perceptions the students held about the CAELD—and its parent, the CAEL—influenced how they valued the diagnostic feedback. The *perceptions of the CAEL/D* code emerged, primarily, from the first set of interviews and, based on the amount of feedback given, chronicled how the students viewed the two tests as either the same or as distinct. It should be noted that I used the appropriate label if the students distinguished between the CAELD and the CAEL; otherwise, I used *CAEL/D* to reflect the two. The students expressed their dissatisfaction with the CAEL/D because of the sustained topic-based test design, and their opinions of how well the three skill sections of the test estimated their ability. Further, the students offered their thoughts on how the respective skill sections should be revised. Overall, the students’ perceptions were largely influenced by their previous experiences with the CAEL and/or other large-scale tests, such as the TOEFL or IELTS.

Three of the students interviewed expressed annoyance with the sustained test topic design of the CAEL/D or suggested that the version they took should have been less science-based. Stephen remarked that, in the seven times he had taken the CAEL, his performance had been largely dependent on the topic: “If the subject is so difficult in the reading, I can’t write. I’m not that good in writing the CAEL test” (Stephen Interview 1). Stephen’s comment implies that he thought his performance on the test was not indicative of his ability or effort, but due to external factors like the test topic. Even though Stephen
indicated that he was not proficient at taking the CAEL, he continued to take the test rather than the IELTS or TOEFL iBT, which are both accepted for placement into the EAP program and admission into the university. Neil recommended that the CAEL/D topics should include more “social topics . . . and don’t do something like biology” (Neil Interview 1). Student dissatisfaction with an integrated test topic is not new and research demonstrates that the validity of the interpretations of students’ overall performance is not compromised by difference in test topic (see Fox, 2004). Fox’s study and the current one vary in terms of both the test users—from EAP program coordinators and university administrators, to teachers and students—and the grain size of the results—the CAEL results used by Fox were broader than the CAELD. The main users for the CAEL are the university admissions officers or the EAP program coordinators; with the CAELD, however, the students are principal test users. As such, their perception of the CAEL and the test topic might have influenced their interpretation of the results. If a student does not fully engage in the test based on the test topic, this disinterest could influence the strengths and weaknesses identified. Therefore, the test topic could be a source of construct-irrelevant variance. Although it is beyond the scope of this study, there may be a need for further investigation into the possible correlation between test topic and students’ identified strengths and weaknesses.

Some of the students clearly articulated how successful they thought the CAEL/D was at measuring their language ability and identifying their strengths and weaknesses. Douglas thought the CAEL/D was a good measure for his listening and writing ability, but was inadequate for estimating his reading ability, because it was too easy: “If you even cannot understand this whole topic or whole article . . . you can still do the test well. It cannot . . . show your real academic learning skill, I think. Reading actually should be more focused on the comprehension, I think” (Douglas Interview 1). Douglas went on to say that the reading section needed a grammar component in addition to vocabulary, because for him, “grammar and vocabulary is the basic of learning English” (Douglas Interview 1). Without an understanding of either component, Douglas suggested that students would not be able to understand the meaning created from the words, nor comprehend the nuances intended from the syntax. Neil agreed
that the reading was too easy. Regardless, he advocated that the CAEL/D was appropriate for identifying strengths and weaknesses. In fact, it was his favourite test, based on the listening and writing: “You need to know what the listening means so you can write the essay . . . and you need to write [the essay] by your own logic. So that’s the good part, I love it” (Neil Interview 1). Neil deliberated on the pros and cons of the CAEL/D and determined it was an ideal test to assess his writing. Part of his consideration for the CAEL/D came from comparison to other large-scale language proficiency tests, which typically did not give as detailed feedback as the CAEL/D. Therefore, his judgment of the CAEL/D was made on a—perhaps unfair—comparison to other tests. Overall, the students thought the CAEL/D was an appropriate measure of their language ability and their strengths and weaknesses. However, some highlighted concerns that the reading underestimated their ability because it was too easy.

Guiding the students’ understanding and thoughts about the CAEL/D were their previous experiences with other English language proficiency tests used for university admissions. When Dennis was asked if he thought the CAELD was a good test for showing his language ability, he replied, “I prefer the TOEFL test. . . . When I was in school we did a lot of practicing the TOEFL test, and I found it helpful for improving the listening, reading comprehension, and different topics” (Dennis Interview 1). Though Dennis did not have an opinion about the suitability of the CAEL as a measure of his language ability, he did indicate his preference for the TOEFL because of his experiences practicing for that test. It would seem that Dennis preferred the TOEFL because of his practiced knowledge of it. It is unknown whether Dennis was referring to the new internet-based test or the older paper-based version, which was based on grammar and thought to be easier. Neil also contrasted the CAEL/D to the TOEFL iBT and stated he thought the CAEL/D was more difficult because it required students to “write deeply” (Neil Interview 1), while the TOEFL iBT focused more on vocabulary. In addition, the students noted differences between the difficulty of the skills sections on the CAEL/D and other tests. For instance, Douglas mentioned that Chinese students would talk about which skills were difficult or easy for the various tests: “I think the IELTS listening and writing is much easier than the CAEL test . . . and the
CAEL test in reading is much easier than IELTS. Many Chinese students talk about this one” (Douglas Interview 1). A number of students seemed to frame their judgment of the CAEL/D based on their previous experiences with other English language proficiency tests, and the perceived difficulty level among the various skill sections of the tests.

Some of the students noted that the topic influenced their performance or recommended that it should be more general. A number of students appraised the CAELD as a language proficiency test and speculated that the reading section might have underrepresented their ability, but commended the writing section as useful for determining their strengths and weaknesses.

**Administration.** The students experienced a range of reactions to taking the CAELD—from anxiety to calmness and curiosity. Each experience formed as a result of the students’ prior encounters with tests and perceptions of how the results might or might not benefit them. Due to the results being low-stakes, the students were given more freedom to come late, attend to their religious commitments, or go home if they felt sick. By missing part of the test, unfortunately, there was less opportunity for the students to demonstrate their full ability. The identified weaknesses and strengths might have been different if they had been able to complete the full test.

The testing situation brought about feelings of anxiety in some of the students. Even though Elaine was a strong student, she revealed her aversion to taking a test under timed conditions. She did remark that she was not nervous because of the limited stakes of the results, but when asked if she would like to try it again, she remarked that “Oh my God, I don’t want to do that anymore. It’s really like the time is going but you know you can do it, but you don’t have the time” (Elaine Interview 1). Elaine provided additional insight into her approach to taking tests. According to her, when there were stakes attached to the results, she would work hard on a test because her “mark is affected by something” (Elaine Interview 1). Specific sections of the CAELD made the students uneasy. The sustained topic also induced nervousness in some students. Dennis stated that he was unfamiliar with the topic and felt
“a little bit nervous” (Dennis Interview 1) as a result. The students who reported being anxious while taking the CAELD did so because of the testing conditions of timing, specific sections, or the topic.

A few students displayed no emotion during the interviews or reflections regarding their CAELD test-taking experience. In the first reflection, Student 10 wrote that he got bored during the reading section due to his lack of sleep the night before. For this student to indicate that he was bored during the reading section suggests that the student was neither tense nor motivated to actively engage in the test. In a testing situation where the results impact university acceptance, students may well push themselves through their tiredness to do their best. In contrast, Don compared his experience of taking the CAELD to a daily task that was not extraordinary: “I’m taking the [CAELD] like I’m having breakfast or anything . . . if it’s in the afternoon, I just take a test for lunch” (Don Interview 1). He also stated that he did not prepare for the test because it did not matter. For these students, the CAELD testing experience was something they needed to do that did not concern them in any way.

In contrast to thoughts of apprehension and indifference, some of the students expressed feeling calm or relaxed during the test. For them, the CAELD feedback was a mechanism for learning about themselves and did not have significant stakes. Kevin described himself as being comfortable in the CAELD testing environment. This relaxed approach was distinct from his experience taking the CAEL, where the results held extremely high-stakes for him: “In the CAEL exam if I don’t get the mark in 18 months they are going to stop me. . . . No, scholarship. Yeah, that’s it” (Kevin Interview 1). He expanded by saying that the CAELD was for his learning: “It let me concentrate on my weaknesses. Like [the teacher] told us that whatever you are weak in . . . you are to improve it. They are the benefits, that’s it” (Kevin, Interview 1). Based on Kevin’s first interview, it seems that he was able to distinguish between the two testing situations of the CAEL and the CAELD because of the stakes attached to the results, even though they are the same test in format, layout, timing, and instructions.

As with Kevin, Laura had taken the CAEL before, but remarked that the diagnostic situation was different. The CAELD represented the opportunity for her to learn about herself: “I was curious about
my progress . . . I wanted to know how much I improved in the English language” (Laura, Interview 1).

To see how she had improved, Laura compared her results on the CAELD to those from the CAEL that she had taken six months prior. In contrast, Stephen skipped over any difficult parts of the CAELD since the assessment did not have any stakes for him. As he said, “I was thinking it’s okay to do anything . . . no matter the score. I was excited to do the reading because I thought I would do well” (Stephen Interview 1). These three students also described the testing situation as comfortable and relaxed. They saw the test as an opportunity to learn about themselves. Though none of them indicated that they enjoyed taking the test, they did display positive emotions about their experience.

Due to religious commitments, two of the students missed part of the reading and listening sections of the make-up test. They were able to return for the writing section. One of the students commented that he felt rushed during the reading because they knew they were going to leave. One student went on to say that he did not feel prepared to write the essay because he forgot what the topic was about and had to re-read some of the readings to remember. For this student, the writing construct changed slightly. The CAEL/D writing is based on the premise that by completing the readings and the listening the student becomes an expert in the subject (Fox, 2004). In this case, the student was only able to complete a portion of the reading and none of the listening, and therefore, did not have the same opportunity as the other students to be as prepared to write the essay. In addition to these two students, another student was late, and one left early due to illness. The low stakes of the situation made for a more relaxed atmosphere, thus giving the students more choice over how much they participated in the test. In the higher stakes CAEL setting, students are not allowed to be late, and they must stay in the room until the test is complete. Most importantly, students often have a vested interest to stay and complete the CAEL, since they paid the fees.

Forty-seven students sat in the same room with the same teacher, took the same test, for the same purpose, but reported experiencing different emotions while taking it. The feelings triggered by the situation ranged from apprehension about the timed conditions and test topic to curiosity in learning
about one’s self. Some of the students missed parts of the CAELD due to the low-stakes setting: by not completing a full CAELD, the interpretations made about their strengths and weaknesses were based on less information than interpretations made regarding their peers.

Table 20

*Students Rank Order of Diagnostic Instructional Activities*

<table>
<thead>
<tr>
<th>Activity</th>
<th>Douglas</th>
<th>Dennis</th>
<th>Stephen</th>
<th>Laura</th>
<th>David</th>
<th>Don</th>
<th>Neil</th>
<th>Elaine</th>
<th>Kevin</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diane’s Teaching</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Targeted Teaching</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Learner Profile</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reflections</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other2</td>
<td>6</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>F</td>
<td>LP</td>
</tr>
</tbody>
</table>

*Note.* Acronyms represent: Library Project (LP), Tracking Sheet (TS), Other courses (OC), Group Discussion (GD), Family (F), and Listening Practice (LPR).

**Experiencing Diagnostic Teaching**

The students had a unique experience with how the results were integrated into instruction. For a detailed description of how the overall mode was calculated, please see Chapter 3. Table 20 depicts the students’ self-reported ranking of the diagnostic instructional activities and components that contributed to their targeted learning. Included in the right-most column is the overall mode of each ranked activity or component. Overall, the highest ranked diagnostic activities according to students were for Diane’s teaching or the targeted teaching. Therefore, either Diane’s teaching or the targeted group teaching was ranked first or second by the students, apart from Elaine, suggesting that some form of teaching was
most important to them and their learning. The rank order of activities in the diagnostic assessment process provided insight into what the students identified as most valuable to the learning.

**The learner profile.** The students’ identified weaknesses were summarized for the students and Diane in the learner profile. Most of the students ranked the learner profile as third or fourth in their ranking of the diagnostic instructional activities and components, or not at all, as was the case with Stephen and Kevin. The students’ perceptions of the learner profile varied: Some saw it as primarily an instructional tool, while others felt it was an aid to help in their writing.

Some of the students remarked that the learner profile was primarily an instructional tool for the teacher, Diane, and not useful for them. Indeed, Kevin and Stephen did not even include the learner profile in their ranking of the diagnostic activities (see Table 20). During the interviews, the students commented on the functionality of the learner profile as a grouping mechanism for the teacher. Dennis commented that the main advantage of the learner profile was that “the teacher can assign the weakness area exactly” (Dennis Interview 1). Only certain sections of the learner profile were identified as useful by the students. Douglas referred to the self-assessment portion of the learner profile as useful for the teacher, but not for students: “I think the [self-assessment] is for teachers. These two things [pointing to the scores and description of weaknesses] are for us. That’s it” (Douglas Interview 1). The students also discussed the learner profile’s usefulness as an instructional tool: While some noted that the teacher was able to use the profile to group them by their weaknesses, others questioned the purpose of the self-assessment portion.

In contrast, Elaine was the only student to rank the profile as most important, because it was the source for describing her strengths and weaknesses: “The learner profile is the main thing right . . . this is the thing we work on” (Elaine, Interview 2). Even though she recognized the importance of the profile as an information source, Elaine admitted that she did not look at the profile beyond the first day. This admission was common across the interview and reflection data, with the exceptions of the students given a balanced profile, those for whom writing was an identified weakness, and Douglas. The students
with an identified weakness in writing—whether they were placed in the writing group or the balanced group—mentioned using the checklist in the profile as a guide and reminder when writing. Douglas reported revisiting the profile at least once a week to keep himself on track for the weaknesses on which he needed to focus.

The students described the features of the learner profile that they liked and made recommendations for how it should be improved. The majority of the students indicated that the comments were the most useful even though all of them recalled looking at the overall scores first. Neil was the only student who said he looked to see the group he was placed in. He wanted to know his main weakness. He then reported looking at the comments. There were three specific recommendations made for how to improve the learner profile. First, Douglas advised for future profiles to have suggestions on how to practice and improve the subskills. Neil recommended that the feedback be given for all of the skill areas and that the students be given the responsibility to continue to improve the skills on their own. Finally, Don identified the need to put the strengths and weaknesses in perspective. He wanted to know that other students had similar challenges, and he was not alone in trying to improve the skills.

The learner profile communicated the CAELD feedback to the students. Overall, they seemed to feel that it was a useful component of the diagnostic assessment process. To improve on its utility as an instructional and learning tool, the students highlighted the need to make it more detailed with comments about all of the skill areas and suggestions on how to practice and suggested the results be put in context by situating the weaknesses and comparing them to other students.

**Perceptions of teaching: Diane’s and the targeted groups**. Just as the students’ perceptions of the CAEL/D were important to their valuing of the feedback, so were their perceptions of how the results were utilized in the course instruction for the intermediate-level EAP course. There were two key components to the course instruction to which the student interviews and reflections referred: Diane’s teaching and the targeted groups.
**Diane’s Teaching.** Diane was responsible for organizing and arranging the feedback, which she also used to identify overall weaknesses of the course. The students articulated a range of opinions regarding the degree to which the feedback was used by Diane in her individual teaching. Neil’s response to why he ranked Diane’s teaching as most important was, “Of course it must be [Diane]. She needs to know the different situations every student has different problems, so if she gets the feedback she can get a better idea of their weaknesses . . . so the students can improve a lot” (Neil Interview 2). This quotation highlights how Neil perceived that Diane used the feedback. In contrast, there were also a few students who commented that Diane’s instruction did not include the CAELD feedback. Douglas stated that Diane did not have enough time to focus on each individual student. “[Diane] needs to follow the outline that is decided before the course begins. I think [Diane] cannot tell the weakest skills of every student. [She] is trying to teach a class and cannot focus on the personal time of students” (Douglas Interview 2). Here Douglas seemed to feel that it was impossible for Diane to incorporate the feedback into teaching. For him, her teaching was predetermined by the course curriculum, determined by the course outline.

Throughout the interviews and reflections, the students identified particular aspects of Diane’s teaching that helped target their learning. The speed-reading tasks were identified as meaningful to the students across the interviews and reflections. Some of the students commented on practicing speed-reading tasks in their studying outside of class. Based on the CAELD feedback, Diane introduced speed-reading tasks into her individual teaching. Don noted that Diane’s feedback on his essay was similar to that of the CAELD feedback; her comments helped confirm his weaknesses in writing. However, most of the students did not perceive any direct links between the CAELD feedback and Diane’s course instruction.

**Targeted Groups.** The students’ quotations revealed that the ability to practice once a week was a central reason for the students’ reported successful use of the CAELD feedback. For instance, David communicated that he would not have used any of the results to modify his studying habits, were it not
for the weekly focused practice on his reading skills. While the targeted groups’ culminating assignment was the portfolio, the students seemed more focused on coming once a week and practicing, than on showcasing their progress in the portfolio assignment. For instance, Don stated that it was useful to “not waste other students’ time” (Don Interview 2). He reflected that it would be disadvantageous for some students whose weaknesses were not the same as the majority. The students, who did not have the common weakness, might have become bored if the emphasis was directed to only those diagnostic subcomponents. Overall, the students seemed to recognize the benefit of the differentiated instruction to specific skill-based weaknesses in the targeted groups.

The collective element of the targeted groups was noted in a number of the students’ interviews. Elaine mentioned that the group discussion helped provide further feedback for her weaknesses in writing as identified by the CAELD, as well as skills not included in her CAELD feedback, on listening and speaking. Don remarked that working together was beneficial because he learned more strategies for targeting his learning.

Though a number of the interview students ranked Diane’s teaching as the most important component of the diagnostic process, their comments nonetheless suggested that the targeted groups had the most direct effect on their learning. The targeted practice served as a motivation to continue to work on their skills. Further, the differentiated approach allowed for the grouping of similar students, which afforded them the opportunity to work on common goals together.

The reflections. The act of reflecting is deeply personal. The students affirmed this notion through their comments in the interviews. The majority of the students reported that the reflections were too prescriptive and mark-based to allow for the reflective process to occur. Though some of the students stated that the reflections were useful, as they reminded them of their weaknesses and helped them articulate a plan on how to improve, others thought that the reflections should not be included in the course.
A few students stated that the reflections were helpful for their learning. Douglas and Elaine had ranked the reflection activity as 3 and 7, respectively. Douglas explained that the reflections helped him to outline a specific plan and link it to the subsequent reflections. Between the reflections, Douglas indicated that he followed a specific plan each week to practice and improve his listening skills. Elaine’s comments in the second interview reflected her low ranking of the reflection. She stated that the in-class reflections “only helped me to remind me everything I guess and that’s all” (Elaine Interview 2). Elaine further explained that she needed to write her own reflections to address the weaknesses for herself. Don felt that it was beneficial to reflect, but did not like the way that marks were attached to the reflections. He stated that he wasn’t able to complete the reflections truthfully since he was writing what the teacher wanted to hear. He recommended that there should be no marks attached, so that students could focus solely on the feedback and thus take the process seriously. According to him, students have to answer the questions honestly to gain the most benefit from the diagnostic assessment process.

Conversely, a number of students stated that the reflections were not at all helpful and recommended that this step should not be included in the program. David declared that the reflections were not helpful for him because they asked him to “to think about how [he] could improve” (David Interview 2). Since he did that already and he knew “everything” (David Interview 2) about himself, he did not see the point in doing the reflections. He was uncertain if the reflections should be removed from the course instruction, since he did not know how the teacher used them. Laura confirmed David’s opinions and also stated that the reflections should be removed from the course instruction.

Though the students recognized the useful process of reflecting and considering how to address their weaknesses or monitor their progress, they stated that the reflections were not a vital piece to include in the course instruction for diagnostic assessment.

**Interpreting and Using the Feedback**

The students were individual in how they reacted, interpreted, and applied the CAELD feedback. Depending on if they agreed with the feedback or not, the students used the feedback differently, as seen
in the *Reactions to and use of feedback* code. Indeed, the students’ perceptions of diagnostic assessment changed from the beginning to the end of the course as seen in *evolving conceptions of diagnostic assessment*. The students’ perceived strengths and weaknesses aligned with the CAELD feedback.

*Figure 3.* Reflection 1 student response counts for if they thought the CAELD feedback reflected their actual language ability.

---

**Reactions to and use of feedback.** The students’ reactions to and use of the CAELD feedback were mixed. In response to Question 1 in the first reflection, students reported that, upon receiving the feedback, they agreed that the results reflected their language ability. However, not all of the students’ qualitative explanations corresponded to their Likert choices, indicating that they did not wholly agree with the CAELD feedback. Further, the students reported contemplating the results at a constant rate throughout the term. *Figure 3* reports the frequency counts to the question, “How much do you think that the diagnostic information presented in the learner profile reflects your actual language ability?” (Reflection 1). More than half of the students (22 of 35) thought that the results conveyed their language ability with responses of *quite a lot* and *completely*. The students’ explanations for their responses differed across and within the Likert choices. The next three subsections describe the students who disagreed, were content, or agreed with the results.
**Disagreed.** All of the students who checked *a little*, and 40% of the students who checked *somewhat*, disagreed with the results. The students asserted that the CAELD feedback over- or under-represented their actual language ability. Most of those who felt the CAELD feedback overestimated their ability, arrived at this conclusion by comparing their performance on other language proficiency tests, such as the IELTS. The students did not indicate if they thought the other language test scores were more accurate in representing their ability. Interestingly, one student explained that he knew the learner profile was not a reflection of his ability since he did not pay attention to the writing, and focused solely on reading when taking the CAELD. Since the test was for diagnostic purposes, the student reported focusing his energies on the reading section, knowing that he needed to target that skill over the others. The students who interpreted the results as underrepresenting their ability provided explanations for their poor performance on the test. These ranged from having had no sleep, being bored during the test, missing part of the test, or not being in the mood. For instance, Student 20 explained that the results did not represent her or his “ability because the diagnostic test it was before the course begin. So I cannot do it well. I was so nervous. I can’t get into that mood and environment. I think it can reflect a little my actual language ability” (Reflection 1).

The student highlighted the feeling of being unprepared for the test and nervous because of the limited class time before the test. Even though the students did not agree with the results, none of them checked *not at all* as a response to the Likert question.

**Content.** A few of the students stated that they were hoping for higher overall or subtest scores, but accepted that the results were accurate and indicative of their strengths and weaknesses. In the first interview, Neil mentioned that he was hoping for “more than 60 but the score is not perfect. But now I know my problem. My problem is writing and use of personal pronouns. I need to learn how to write papers using the reporting verbs” (Neil, Interview 1).

**Agreed.** More than half of the students (22 of 35) responded *quite a lot* and *completely*, conveying their belief that the results accurately represented their language ability. These students
remarked that the weaknesses reported in their CAELD feedback reflected the areas where they needed to improve. Finally, some of the students commented on the usefulness of the results. For instance, Student 36 perceived that the CAELD listening was more academic in nature, compared to the IELTS, “and the comments have given me some good advice. I really need to improve my listening” (Reflection 1). The findings regarding the students’ reactions to the CAELD feedback highlight the need to draw on multiple data sources from students in order to gain a rich understanding of how truly they perceive their language ability.

**Perceived strengths and weaknesses.** Sixty-six codes pertaining to the students’ perceived strengths and weaknesses emerged from the student interview and reflection data. Less than one quarter corresponded to the students’ perceived strengths, while the remaining focused on their weaknesses. Tables 21 and 22 list the students’ perceived strengths and weaknesses in relation to the CAELD diagnostic subcomponents.

Table 21

*Students’ Perceived Strengths (13 total)*

<table>
<thead>
<tr>
<th>Reading (3)</th>
<th>Listening (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• In general (2)</td>
<td>• In general (2)</td>
</tr>
<tr>
<td>• Locating supporting details*</td>
<td>• General listening (based on IELTS score)</td>
</tr>
<tr>
<td>Writing (5)</td>
<td>Speaking (2)</td>
</tr>
<tr>
<td>• In general</td>
<td>• In general (2)</td>
</tr>
<tr>
<td>• Organization of paragraphs into logical sequence*</td>
<td></td>
</tr>
<tr>
<td>• Coherence</td>
<td></td>
</tr>
<tr>
<td>• Grammar</td>
<td></td>
</tr>
<tr>
<td>• Sometimes strong in writing the CAEL essay</td>
<td></td>
</tr>
</tbody>
</table>

Note: The asterisk* denotes the strength as corresponding with the CAELD diagnostic subcomponents.

Similar numbers of quotations were coded for students’ perceived weaknesses of reading and writing. However, the level of detail provided by the students was less for reading weaknesses than for writing. When describing their weaknesses in reading, the students often did not specify any particular aspect where they felt they were weaker. For instance, Laura stated that she was most concerned about
her reading ability: “It’s my weakness reading. I was curious about everything, how I can I understand the reading, how to use strategies” (Laura, Interview 1). Since Laura was interested in improving all aspects of her reading skill, she responded that her weakness was reading in general. In contrast, other students might have described their weakness in reading generally because of how they interpreted the results from the learner profile. Kevin described his experience of receiving the learner profile and the details that he focused on: “[The learner profile] gave me the information that reading was my problem. I have to improve it. My mark was 50, but the problem was reading. It was 40” (Kevin, Interview 1). For Kevin, the CAEL band scores provided the evidence of his strengths and weaknesses. Kevin attributed his weak reading skills to the fact that he did not like to read in his first language, Arabic. Thus he seemed to draw on both the CAELD feedback and his own knowledge about himself in forming his idea of being weak in reading.

Table 22

Students’ Perceived Weaknesses (54 total)

<table>
<thead>
<tr>
<th>Reading (21)</th>
<th>Listening (10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• In general (10)</td>
<td>• In general (4)</td>
</tr>
<tr>
<td>• Reading speed (5)</td>
<td>• Academic listening (3)</td>
</tr>
<tr>
<td>• Reading every word</td>
<td>• Catching specific information* (2)</td>
</tr>
<tr>
<td>• Understanding the gist* (2)</td>
<td>• Difficulties concentrating</td>
</tr>
<tr>
<td>• Identifying the main idea* (2)</td>
<td></td>
</tr>
<tr>
<td>• Locating supporting detail*</td>
<td></td>
</tr>
<tr>
<td>Writing (19)</td>
<td>Vocabulary (5)</td>
</tr>
<tr>
<td>• In general (4)</td>
<td>• Limited vocabulary (3)</td>
</tr>
<tr>
<td>• Word category choices and use are weak* (2)</td>
<td>• Spelling</td>
</tr>
<tr>
<td>• Coherence (2)</td>
<td>• Guessing the meaning from the context</td>
</tr>
<tr>
<td>• Grammar (3)</td>
<td></td>
</tr>
<tr>
<td>• Structural fragments*</td>
<td></td>
</tr>
<tr>
<td>• Informal (colloquial) vocabulary* (2)</td>
<td></td>
</tr>
<tr>
<td>• Subject verb agreement* (2)</td>
<td></td>
</tr>
<tr>
<td>• Inappropriate translations</td>
<td></td>
</tr>
<tr>
<td>• 1st paragraph does not give the reader an overview of the writing to follow*</td>
<td></td>
</tr>
<tr>
<td>• Body of writing is not divided into logical paragraphs and sections*</td>
<td>• Time management</td>
</tr>
</tbody>
</table>

Note: The asterisk* denotes the weakness as corresponding with the CAELD diagnostic subcomponents.
Thinking about and using the feedback. The students’ consideration of the CAELD feedback varied over time and by context. Figure 4 displays the students’ responses to the question, “How often have you thought about the diagnostic feedback (learner profile) when studying in the [Intermediate-level EAP course]?” for Reflections 2 and 3. There was no real change between reflection 2 and 3 for how often the students reported thinking about the CAELD feedback in class; 29 students reported thinking about the feedback either once a week or more than once a week for both reflections. The trend for the increase in the once a week category is likely due to the weekly, targeted groups that started the same week Reflection 2 was completed.

Figure 4. Reflection 2 and 3 student responses indicating if they thought about the CAELD feedback during the Intermediate-level EAP course.

![Bar graph showing student responses to thinking about the CAELD feedback.](image)

Fewer students reported thinking about the results outside of class. Figure 5 reports on student responses to the question, “How often have you thought about the diagnostic feedback (learner profile) when studying on your own and outside of class?” for Reflections 2 and 3. The student response rate, which was relatively constant between Reflections 2 and 3, indicated that even though more students thought about the results during class, this did not transfer to their studying outside of class. It would be expected that the frequency in which the students thought about the CAELD feedback would decrease as
the course progressed. Having the weekly, targeted groups might have influenced the students to consider the results more than if they did not have these groups.

Figure 5. Reflection 2 and 3 student responses indicating if they thought about the CAELD feedback on their own and outside of class.

![Figure 5](image)

**Evolving conceptions of diagnostic assessment.** The students’ definitions of *diagnostic assessment* evolved as the course progressed. Though most of the students stated that they thought the diagnostic assessment sounded like a good idea, with the exception of Neil, they were unsure what to expect or thought its primary use was for Diane’s teaching. Neil was one of the few students who indicated that he saw diagnostic assessment as a way for him to improve his language skills.

In fact, at the beginning of the course, the students defined the concept based on their previous experiences with other courses, other assessments or self-assessments, or from their expectations of what diagnostic assessment would look like. The students’ experiences with the intermediate-level EAP course seemed to further influence their definition of diagnostic assessment, as seen in their comments at the end of the course. Table 23 depicts this shift in how four of the students conceptualized diagnostic assessment at the beginning and at the end of the course.
Table 23

Students’ Perceptions of Diagnostic Assessment at the Beginning and End of the Course

<table>
<thead>
<tr>
<th>Student</th>
<th>Beginning of the course</th>
<th>End of the course</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elaine</td>
<td>C: Have you experienced diagnostic assessment before?</td>
<td>C: What is the goal of diagnostic assessment for teachers?</td>
</tr>
<tr>
<td></td>
<td>S: Depending on my marks I would be placed in different groups. If I get a bad mark the teacher will talk to me you have to study this, this, this, and come back again (Elaine Interview 1).</td>
<td>S: Make their students understand that this result is very important for you guys and to work on it and then you can improve your language skills. (Elaine Interview 2)</td>
</tr>
<tr>
<td></td>
<td>C: Did you enjoy it?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S: No, a lot of studying. (Elaine, Interview 1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Subcode: Evaluative</td>
<td>Subcode: Formative</td>
</tr>
<tr>
<td>Dennis</td>
<td>C: Okay, and what about the term diagnostic assessment, have you ever heard of that term before?</td>
<td>C: Okay. So now that you’ve experienced this course with diagnostic assessment, what advice would you give a student in the course?</td>
</tr>
<tr>
<td></td>
<td>S: Yes I took twice, 2009 in May and one in July.</td>
<td>S: To work hard, harder than like they think it’s hard in the targeted group because they have to improve this weakness more than anything else, like focus more on this. It’s like it’s a really good chance for them to improve their weakness. (Dennis Interview 2)</td>
</tr>
<tr>
<td></td>
<td>C: Okay, so what you mean is that you took the CAEL before?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>S: Yes the CAEL. So I think it was similar the diagnostic, it’s the same I think. (Dennis Interview 1)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Subcode: Evaluative</td>
<td>Subcode: Formative</td>
</tr>
<tr>
<td>David</td>
<td>C: Have you experienced diagnostic assessment before?</td>
<td>C: Okay. So I just have one question, what then is the goal of diagnostic assessment?</td>
</tr>
<tr>
<td></td>
<td>S: Yeah, it’s like, it’s my first time that I have to take a test before the course . . . to know the weakness . . . to fix it. That’s why I think I have to take this test. (David Interview 1)</td>
<td>S: To let the student know the weakness and where they are now in their English skills and how to be better and how to improve it, to know like where is their place like where are you now in English. (David Interview 2)</td>
</tr>
<tr>
<td></td>
<td>Subcode: Evaluative</td>
<td>Subcode: Formative</td>
</tr>
</tbody>
</table>

When discussing diagnostic assessment, the students described it as evaluative of their learning or similar to a placement test. For instance, Elaine described her previous diagnostic experiences from Grades 7 to 12 as evaluative and mark-based (see Table 23). The subcodes of *evaluative* and *formative* were used to note changes in the students’ evolving conceptions of diagnostic assessment. It would seem...
that Elaine’s previous diagnostic experiences occurred at the end of her learning rather than at the beginning, not giving her the chance to maximize the learning opportunity. Elaine’s explanation of the diagnostic assessments she had taken before seemed punitive in contrast to her articulation of the concept at the end of the intermediate-level EAP course, which was based in her experience in the course. In particular, she viewed the teacher’s role differently—as that of a motivator, rather than an enforcer. Don and Stephen also likened diagnostic assessment to higher-stakes evaluations, such as midterm or placement tests, at the beginning of the course. Meanwhile, Laura’s conceptualization of diagnostic assessment as looking at strengths and weaknesses was based on her practice of self-assessing her own ability. Similar to Elaine’s verbalization of diagnostic assessment, Laura grounded her description at the end of the course based on her experience in the course.

Dennis seemed to think that the diagnostic assessment was the same as the CAEL, since the test used was the CAELD. He attributed the term to the test, rather than the outcome and influence on teaching. At the end of the course, he stressed the need for students to take the assessment seriously because the results would impact their opportunity for learning. David’s definition of diagnostic assessment was similar to the one given by Diane at the beginning of the course—to better understand the students’ weaknesses. However, earlier in the first interview, David indicated that he did not like having the diagnostic assessment at the beginning of the course because he considered test taking to be evaluative. Further to the point, he remarked he was unsure why the CAELD was given at the beginning of the course, as students had not learned anything yet. He assumed that test taking could only occur after teaching, so it would be fair to measure his learning. His definition of diagnostic assessment was similar to the one given at the beginning of the course but more descriptive.

Whether the students viewed diagnostic assessment as a judgment of their ability or a support to their learning varied, based largely on their previous experiences with assessment. After their experiences in the course, the students expanded their notion of diagnostic assessment: They came to see
diagnostic assessment as a nonjudgmental tool not only for the teacher, but also for the purpose of their own learning.

**Learning goals.** Across the interview data, there were matches and mismatches between the CAELD feedback and the students’ learning goals. Further, some of the students looked to the CAELD for their language-learning goals.

For instance, there was a clear match between David’s learning goal and his weakness. David stated that if reading had not been labelled as his weakness, he would not have focused or cared to try to improve: “I have to agreed this is my weakness. . . . I know reading is my weakness, so I want to improve it and I wasn’t surprised by the result of the test” (David Interview 2). There was a potential mismatch identified between Dennis’ learning goals and his weakness in writing, which led to his placement into a writing targeted group. He admitted that his weakness was probably writing, but he did not need it as a master’s electrical engineering student. He felt he wanted to improve his reading because that was the skill he used the most and found the most challenging in his content courses: “We don’t write essays, maybe once: sometimes we don’t do any essays. . . . We should focus on reading, this may be the most challenging [for me]” (Dennis Interview 1). Similar sentiments were expressed by Laura. In contrast to David and Dennis, who internally identified learning goals, Douglas set his goals based on the CAELD feedback. For him, the CAELD told him “what [he needed] to learn in the future” (Douglas Interview 1). He seemed to completely trust the CAELD feedback to inform him of what his language-learning goals should be. Regardless of the source of the learning goal, the students who made connections between the CAELD feedback and their own learning goals seemed to express more interest in studying to improve their targeted weaknesses.

The students made multiple connections between the CAELD feedback and how and why they sought to improve their targeted weaknesses. For some, there was a match between the language goals they had already set for themselves and the CAELD feedback; others saw a mismatch between their future learning needs and the focus of the instruction suggested by the results, as was the case with
Dennis. The connections that the students made throughout the process were vital to the extent that the students used the feedback in how they approached their learning.

Table 24

Approaches Students Identified Helping Them Target Their Weaknesses (Review and Learn; Practice)

<table>
<thead>
<tr>
<th>Review and Learn</th>
<th>Practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Vocabulary (8)</td>
<td>• Reading (12)</td>
</tr>
<tr>
<td>o Read books/newspapers</td>
<td>o Read books/research</td>
</tr>
<tr>
<td>o Study the suffix endings and prefix beginnings of words</td>
<td>o Find interesting books to read</td>
</tr>
<tr>
<td>o Learn synonyms</td>
<td>o Timed reading</td>
</tr>
<tr>
<td>o Forms and functions</td>
<td>• Listening (7)</td>
</tr>
<tr>
<td>o Practice pronunciation</td>
<td>o Watch TV/movies to understand fast-paced speech</td>
</tr>
<tr>
<td>o Focus on high frequency words</td>
<td>o Listen to English music/news</td>
</tr>
<tr>
<td>• Grammar (2)</td>
<td>o IELTS practice listening exercises</td>
</tr>
<tr>
<td>o Word order/structure</td>
<td>o Vocabulary exercises to improve listening skill</td>
</tr>
<tr>
<td>• Writing (6)</td>
<td>o Practice note-taking skills in lectures</td>
</tr>
<tr>
<td>o Reporting verbs and reference statements for writing academic papers</td>
<td>• Vocabulary (5)</td>
</tr>
<tr>
<td>o Transition exercises</td>
<td>o Use the subtitles to practice knowledge of vocabulary</td>
</tr>
<tr>
<td>o Read to observe how writers write</td>
<td>o When reading, try to understand the meaning of words from context</td>
</tr>
<tr>
<td></td>
<td>o Write down keywords when listening to news</td>
</tr>
<tr>
<td></td>
<td>o Recite new vocabulary words every day and spell them without looking</td>
</tr>
<tr>
<td></td>
<td>• Writing (1)</td>
</tr>
<tr>
<td></td>
<td>o Write summaries of readings</td>
</tr>
<tr>
<td></td>
<td>• Grammar (1)</td>
</tr>
<tr>
<td></td>
<td>o Buy a book and practice one chapter a week</td>
</tr>
</tbody>
</table>

Approaches taken to target learning. The students employed four types of approaches to help target their weaknesses, whether focusing on broad skills or on the diagnostic subcomponents identified by the CAELD feedback: review and learn, practice, strategy, and seek advice and resources. When the students recognized that they needed to acquire more knowledge about a skill before they could practice it, the approaches were classified as review and learn. Similar to a muscle that needs exercise to become stronger, the students recognized that they had to practice the skills and diagnostic subcomponents in order to improve. The strategies employed by the students were more specific and were employed in the specific context in which they needed to use the skill or diagnostic subcomponent. To help them learn
more about the skill or diagnostic subcomponent, the students sought out help from other people or resources, such as grammar books, to gain new knowledge or strategies.

Table 25

*Approaches Students Identified Helping Them Target Their Weaknesses (Strategy: Engage in Further Learning)*

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Engage in further learning (seek advice and resources)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writing (12)</td>
<td></td>
</tr>
<tr>
<td>• Use writing checklist from the learner profile to help write essays/research project</td>
<td></td>
</tr>
<tr>
<td>• To remember the typical grammar mistakes made</td>
<td></td>
</tr>
<tr>
<td>• As a guide for writing</td>
<td></td>
</tr>
<tr>
<td>• When writing essays, remember correct word order</td>
<td></td>
</tr>
<tr>
<td>• To form complete sentences: Think about what to write in first language and then translate.</td>
<td></td>
</tr>
<tr>
<td>• I will list a writing plan that is specific, rather than general. You do one by one.</td>
<td></td>
</tr>
<tr>
<td>• Include academic words in vocabulary by using dictionary</td>
<td></td>
</tr>
<tr>
<td>• Colour code mistakes when writing</td>
<td></td>
</tr>
<tr>
<td>• Ignore new words and look at the surrounding words</td>
<td></td>
</tr>
<tr>
<td>• Avoid using personal (non-academic) words. Use reporting words instead.</td>
<td></td>
</tr>
<tr>
<td>• Notice if text is cohesive and not choppy</td>
<td></td>
</tr>
<tr>
<td>• Focus on or agree what the question/prompt is asking</td>
<td></td>
</tr>
<tr>
<td>Reading (4)</td>
<td></td>
</tr>
<tr>
<td>• Focus on the first and last sentence of paragraphs to guess the meaning</td>
<td></td>
</tr>
<tr>
<td>• Identify only the information related to the questions</td>
<td></td>
</tr>
<tr>
<td>• Put brackets around new words</td>
<td></td>
</tr>
<tr>
<td>Listening (1)</td>
<td></td>
</tr>
<tr>
<td>• Watching movies: The first time try to understand the main idea. The second time try to “catch” the keywords.</td>
<td></td>
</tr>
<tr>
<td>Writing (6)</td>
<td></td>
</tr>
<tr>
<td>• See help to improve essays</td>
<td></td>
</tr>
<tr>
<td>• Buy a book on how to write essays and properly use transitions (move to resources)</td>
<td></td>
</tr>
<tr>
<td>All skills (2)</td>
<td></td>
</tr>
<tr>
<td>• Take the CAELD again to learn about progress</td>
<td></td>
</tr>
<tr>
<td>Reading (1)</td>
<td></td>
</tr>
<tr>
<td>• Read books on how to start introductions (resources)</td>
<td></td>
</tr>
</tbody>
</table>

For each of the subcodes, there was a dominant skill targeted in the approach. For example, in *review and learn*, vocabulary was the main skill on which students focused. When practicing, the students primarily targeted their reading skills, both comprehension and speed. In order to avoid mistakes when writing, the students highlighted numerous strategies. Interestingly, a few indicated that
they would use the checklist in their learner profile to aid them in their writing. Once again, writing was the prominent skill targeted when the students sought out external help from support services and resource materials. Most of the approaches described by the students were specific and targeted a particular diagnostic subcomponent, apart from practice, which focused broadly on the four skills. Though vocabulary was not a separate skill on the learner profile, it was targeted specifically by the teacher in the course (see Chapter 5). Tables 24 and 25 lists the approaches the students reported using in the interviews and reflections. The frequency that the students referred to the approach taken is represented in brackets next to the main skill addressed. To help improve their learning, the students employed a variety of approaches (review and learn, practice, strategies, and seek advice and resources). The main skills targeted were vocabulary, reading, and writing.

**Discussion**

Learning an additional language is a complex process. Beyond the cognitive processes and acquisition of grammatical structures, the learning is embedded within a course influenced by a particular pedagogic practice. Adding to the complexity are the students’ personalities and underlying motivational goals for learning academic English.

**Facilitating Question 3.1:** How did the students use the diagnostic feedback to benefit their learning?

The students reported thinking about the results inside and outside of the intermediate-level EAP course at a relatively constant rate throughout the term. They reported using a range of approaches to target their weaknesses, as identified by the CAELD feedback. Most of the approaches taken focused specifically on the subdiagnostic components listed on their learner profile. Another dimension of the students’ use of the CAELD feedback depended on the degree of autonomy with which the students approached their learning. Students in this study benefited from instruction based on the CAELD feedback, a finding mirrored in earlier research on the TOEFL iBT for diagnostic purposes (Jang, 2005, 2009).
Facilitating Question 3.2: How did the students interpret the CAELD feedback?

The students’ interpretation of the CAELD feedback was influenced by their perceptions of the CAEL and CAELD in relation to other large-scale proficiency tests used primarily for university admissions purposes. The students seemed to interpret the results first, broadly, by skill and then in more detail, using information related to the diagnostic subcomponents. When articulating their strengths and weaknesses, the students’ comments were primarily general for all skills except writing.

Facilitating Question 3.3: How did the diagnostic feedback map onto the students’ learning goals?

There was evidence of a match between the students’ learning goals and the CAELD feedback. This match was partially based on whether the students’ goals were related to their language learning in general or to their academic studies. The students who expressed language-learning goals found that their goals matched the CAELD feedback.

With students who articulated language-learning goals specific to their academic goals—such as being an engineering student—there was more potential for a mismatch. This was the situation with Dennis, whose own academic language goal was to improve reading, even though he was identified as being weak in writing. Dennis’ disinterest in writing, grounded in his perceived needs as an engineering student, was similar to Ben’s, an engineering student in Leki’s (2007) case descriptions of English language learners in university. According to Leki, Ben did not value writing instruction as much as the other skills taught in his language support courses because, in his experience, he did not write in his courses as much as he used the other skills. Cheng and Fox (2009) argue that it is important to align EAP instruction with students’ perceived needs to maximize students’ participation and learning. Therefore, to best align the use of diagnostic assessment with students’ learning goals, the results and decisions made could be tailored to the students’ academic language goals, rather than only to their identified weaknesses.
Some of the students seemed to base their language-learning goals on the CAELD feedback. When the students used the results to solely dictate their language-learning goals and possibly influence their identity of themselves as language learners, the CAELD feedback was potentially higher stakes. This scenario is in contrast to the situation of students who already have a sense of their strengths and weaknesses. These students demonstrated a range of learner autonomy; for them, the CAELD feedback might or might not have confirmed their sense of their own abilities. Some of the students needed explicit explanations from the teacher on how to connect their results to their learning, while others illustrated a more mature approach to learning and made the connections themselves.

**Facilitating Question 3.4:** How did the diagnostic feedback map onto the students’ values about learning and assessment?

Initially, because of some students’ previous assessment experiences, there was no alignment between their values of learning and the CAELD feedback. Their experiences influenced their perceptions and definitions of diagnostic assessment. The initial notions of diagnostic assessment that the students held seemed largely evaluative. Some seemed confused by having a test at the beginning of a course. For these students then, there was an initial mismatch between the students’ values of learning and instruction and the CAELD feedback. After experiencing the course instruction and how the CAELD feedback was used, a better match developed between the students’ values of learning and the CAELD feedback. The students clearly placed importance on teaching and instruction, as evidenced by the students’ ranking of the diagnostic assessment activities.

**Facilitating Questions 3.5:** What was the students’ test-taking experience of the CAELD?

Even though the students took the same test, in the same room, with the same instructions, their experiences varied. Some of them were unable to distinguish the lower-stakes CAELD assessment from the CAEL, a high-stakes language-proficiency test. These students’ inability to separate the different purposes was likely based on their previous experiences taking the CAEL for university admissions or EAP program-placement purposes. Another possibility for the expressed discomfort could be the
standardized test-like setting of the CAELD. This anxious view of the CAELD administration contrasted with other students’ keen interest and enthusiasm for taking the test, which they viewed as a learning opportunity. Between the two extremes were the students who displayed apathy towards their CAELD test-taking experiences.

These five facilitating questions inform the research question asked in Phase 3:

**Research Question 3:** What are the learning conditions and students’ perceptions that enable the students to use the CAELD and its results to support their learning?

A key, mediating factor in the entire process was the learning environment facilitated by the teacher: her commitment, previous experience, and her competence in diagnosing her own students. In addition, the targeted groups were also critical to the students’ use of the CAELD results. Support for the importance of the learning context was seen in the students’ evolving conceptions of diagnostic assessment: the teaching and learning context played a significant role in how the students conceptualized diagnostic assessment and the potential the results could have for their learning.

The students who benefited the most from the CAELD were students who were most likely motivated to improve their language skills. As such, the value that the students had for the process and the assessment tool was vital to its success. This value was enhanced through a perceived match between the CAELD feedback and their own self-assessment and short- and long-term language goals. The students whose assessment results matched their goals, found the course to be beneficial. For those few who did not see a connection between the feedback and their language goals, the test was perhaps not as valuable. This variance in students’ perceptions of the initial usefulness of the CAELD and the diagnostic process suggests the teacher needs to emphasize the formative nature of the assessment, and explicitly explain how the results will be used. To strengthen the ability of an assessment intended for making instructional decisions or helping students improve their language skills, the students needed to be included to a greater extent in the decision-making process and implementation of such assessments.
Chapter 7: Assessment Use Argument

This chapter integrates the findings from the study’s three overarching research questions, which together examine the extent to which the Canadian Academic English Language Diagnostic (CAELD) assessment is appropriate for diagnostic use in an English for Academic Purposes (EAP) program. This examination is based on the perspectives of two raters, Isobel and Catherine; one teacher, Diane; and Diane’s 47 students. Nine of these students were interviewed (pseudonyms: Dennis, Douglas, Neil, Elaine, Laura, Stephen, David, Don, and Kevin). This chapter applies the Assessment Use Argument (AUA) framework to examine the findings from this study’s three results chapters (i.e., Chapters 4 to 6). My argument in this chapter also links to previous research on the CAEL and CAELD, so as to present readers with an overall, transparent appraisal of the test.

The AUA for the CAELD

In Chapter 2, I outlined the AUA for the CAELD and the four assessment claims: (a) beneficial consequences, (b) decisions made, (c) interpretations, and (d) assessment records. These claims represent a chain of inferences needed about a student’s test performance. To assess the strength of the argument, I begin with assessment records (Claim D) and follow through to beneficial consequences (Claim A). Table 26 summarizes the assessment claims and the supporting warrants for each category that emerged from this research.

I present the evidence as a series of tables for each assessment claim or warrant category. In each table, I contrast the two evidence categories as backing and rebuttal for each warrant. The backing or rebuttals are derived from the three results chapters as bullet points. This comparison of the two evidence types highlights both the strengths of the assessment claim and the weaknesses, i.e., those areas that require future research of the CAELD use to further enhance the argument. In some instances, I have identified possible rebuttals that are not based on data derived from this current study. These possible rebuttals are areas that were identified as a concern but lacked data to support them from this
study. The evidence forming the backing, rebuttals, or possible rebuttals are referenced by source, to Chapters 4, 5, 6, and to previous research on the CAEL.

Table 26

Assessment Claims for the AUA for the CAELD

<table>
<thead>
<tr>
<th>Assessment Claim</th>
<th>Warrants (Number)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Beneficial consequences</td>
<td>• Beneficial (5)</td>
</tr>
<tr>
<td>B. Decisions made</td>
<td>• Value sensitivity (2)</td>
</tr>
<tr>
<td></td>
<td>• Equitable decision-making (2)</td>
</tr>
<tr>
<td>C. Interpretations</td>
<td>• Meaningful (6)</td>
</tr>
<tr>
<td></td>
<td>• Impartial (5)</td>
</tr>
<tr>
<td></td>
<td>• Generalizable (2)</td>
</tr>
<tr>
<td></td>
<td>• Relevant (3)</td>
</tr>
<tr>
<td></td>
<td>• Sufficient (1)</td>
</tr>
<tr>
<td>D. Assessment records</td>
<td>• Consistent (1)</td>
</tr>
</tbody>
</table>

Assessment records. During the first stage in the assessment process, the students took the test and completed the assessment records. The learner profiles were subsequently created. Assessment Claim D stated, “The scores from the CAELD are consistent across different forms and administrations of the test, and across students from different test-taker groups and academic backgrounds.”

There was only one warrant (D.1) articulated for the assessment claim regarding the consistency of the assessment records (see Table 27). The previous CAEL validation research (CAEL, 2008; Fox, 2004, 2009; Fox, Pychyl, & Zumbo, 1993), the benchmarking session between the teacher and myself as part of the current study, and the final review of profiles by the teacher all support the claim that the assessment records were consistent across the students. In addition, the CAELD was scored under conditions similar to those for the CAEL, suggesting that the reliability information should hold true for the CAELD.

Some rebuttals call into question whether or not the reliability information is, in fact, applicable to the additional diagnostic purpose. The raters’ awareness of the additional, low-stakes diagnostic purpose might have led them to be more lenient or severe in their scoring of the essays. If so, the band scores reported for the students’ CAELD writing would be lower or higher than their true score.

165
This concern for the reliability of the rating suggests that a benchmark session specifically related to the CAELD for diagnostic purposes should be included in the scoring process. Alternatively, the raters could score the essays blind, unaware of the tests’ purpose, i.e., placement, admissions, or diagnostic, which would reduce the likelihood of any bias in the rating.

Another source of systematic error might have occurred during the creation of the learning profiles.

Table 27

Assessment Records Warrant (D.1) for Assessment Claim D

<table>
<thead>
<tr>
<th>Warrant</th>
<th>Backing</th>
<th>Rebuttals</th>
</tr>
</thead>
</table>
| Warrant D.1: The assessment records are consistent across different groups of students (consistency). | • The CAEL has high reliability (CAEL, 2008).  
• The teacher and the author had a benchmarking session for five CAELD learner profiles (Chapter 4).  
• The teacher reviewed the learner profiles before making the final decision about placing students into the targeted groups (Chapter 5).  
• Possible: The students’ self-assessment served as an additional source of evidence for the creation of the learner profiles (Chapter 5). | • Reliability information was not collected for the CAELD.  
• Possible*: The raters’ perceptions of the use of the CAELD and the students’ needs might have influenced their scoring of the CAELD (Chapter 4).  
• Possible: When the teacher encountered a CAELD learner profile she questioned, she sought additional information to inform her decision; however, the learner profiles were also created by the researcher (myself), who did not have access to the additional information (Chapter 5).  
• Possible: The two-stage scoring process likely introduced random error (Chapter 5). |

*Note: The word possible denotes the possible rebuttals identified. Please see the description about possible rebuttals at the beginning of this section, The AUA for the CAELD.

Even though the teacher and I completed a benchmarking session to establish consistent scoring practices, disparity in our knowledge about the students’ language proficiency might have caused error to be introduced. Further, error could have been introduced because of the two-stage marking process: (a) the scoring of essays and question booklets by CAEL raters and (b) the creation of the learner profiles (teacher and students) by the teacher and the researcher. Grounded in Smith’s (2003) idea of...
reliability for classroom-based assessments through sufficiency of information, the students’ self-assessments might have diffused the degree of inconsistencies generated at either of the two stages. Alternatively, they might have increased the error when viewed from a large-scale assessment perspective since there were no formal guidelines on how to incorporate the self-assessment data. Nevertheless, to reduce the potential for error, the two-stage process should be streamlined to a one-stage process, or use the information from the two stages to support each other to increase the validity of the assessment. More research is needed that investigates whether or not the reliability of the raters’ scoring is influenced by knowledge of the additional diagnostic purpose, and to what extent error is introduced through the two-stage scoring process.

**Interpretations.** Forming interpretations about the students’ test performance was composed during the second stage in the assessment process. To this end, Assessment Claim C stated:

The interpretations about students’ strengths and weaknesses in reading, listening, and writing are meaningful to the course syllabus and the teaching and learning activities in the EAP course, impartial to all groups of test takers, generalizable to subsequent learning tasks used in introductory level university courses, and relevant to and sufficient for the formative decisions made by the teacher and students.

The interpretations assessment claim was the most complex to construct since it required the integration of the 17 warrants elaborating the claim: meaningful (6), impartial (5), generalization (2), relevant (3), and sufficiency (1). With so many warrants articulated for the interpretations assessment claim, I discuss each of the warrant categories individually.

**Meaningfulness.** There were six warrants (C.1 – C.6), articulated in relation to the meaningfulness of the CAELD construct, for teaching and learning within the intermediate-level EAP course (see Tables 28, 29, and 30). Specifically, the warrants correspond to how the Target Language Use (TLU) domain and construct was defined, the need for the CAELD administration to facilitate the
highest performance possible from the students, the meaningful communication of the construct to the stakeholders, and in turn, their comprehension of it.

Table 28

Meaningfulness Warrants (C.1–C.2) for Assessment Claim C

<table>
<thead>
<tr>
<th>Warrants</th>
<th>Backing</th>
<th>Rebuttals</th>
</tr>
</thead>
</table>
| **Warrant C.1:** The definition of the construct, based on academic reading, listening, and writing an argumentative essay, clearly distinguishes this construct from other constructs of language assessment. The construct of the diagnostic subcomponents clearly distinguishes them from other diagnostic subcomponents (construct and TLU domain definition). | • The construct definition was based on a needs analysis of first-year university courses (Fox et al., 1993); and a retrofit study linking the test specifications to diagnostic subcomponents, supported by the research literature (Fox, 2009).  
• The strengths and weaknesses identified by raters were more detailed at the higher score levels (i.e., ≥40; Chapter 4).  
• Reading speed was identified as useful by the teacher and students (Chapters 5 and 6). | • The higher scored essays elicited more detailed observations from raters than the lower scored essays (Chapter 4).  
• Reading speed was identified as a targeted skill by the teacher, but was not exclusive from the other reading subcomponents (Chapter 5).  
• Vocabulary was a predominant skill the teacher targeted in her lessons (Chapter 5).  
• Some students identified the need for sections on grammar and vocabulary (Chapter 6).  
• Possible: The reading subcomponents might only be meaningful for students at a particular proficiency level (Chapter 5). |
| **Warrant C.2:** The CAELD engages the student in academic reading, listening to, and writing an argumentative essay (accuracy of inferences). | • The raters’ discussion of students’ strengths and weaknesses represented the construct articulated in the scoring guide (Chapter 4).  
• The teacher indicated that the writing sections of the CAELD were appropriate measures for identifying students’ strengths and weaknesses (Chapter 5).  
• Some students perceived the CAELD as an appropriate tool for eliciting academic English skills (writing and listening) in order to identify strengths and weaknesses (Chapter 6). | • The teacher commented that the listening and reading sections of the CAELD were too difficult, which made identifying students’ strengths and weaknesses problematic (Chapter 5).  
• Most students commented that the reading section was not challenging enough to identify diagnostic subcomponents (Chapter 6). |

The first warrant (C.1) ensures the construct definition is grounded in a frame of reference, such as needs analysis, research, or theory (Bachman & Palmer, 2010). The second part of the warrant qualifies the construct as distinguishable from other similar abilities in the TLU domain (e.g., listening
to a lecture versus listening to friends talk about a lecture). There is conflicting evidence (backing and rebuttals) that questions the current definition of construct for the intermediate-level EAP course. The construct and identification of the TLU domain were based on Fox et al.’s (1993) needs analysis, which established the CAEL construct, and on Fox’s (2009) more recent retrofit study for diagnostic subcomponents. Nonetheless, rebuttal data suggest that the CAELD TLU domain is not specific enough for the intermediate-level EAP course: The construct underrepresents the targeted skills for the stakeholders’ needs. However, evidence gained from the three chapters demonstrates that the CAELD essay elicits students’ strengths and weaknesses for writing academic English, and therefore supports the construct.

However, backing and rebuttal data from Chapter 4 suggests that the type and number of strengths and weaknesses that can be drawn from the essays, varies with the students’ proficiency level. In other words, it was more challenging for Isobel and Catherine to identify strengths and weaknesses in the low-level essays (i.e., ≤30) than in the higher levels (i.e., ≥40). In the lower level essays, the diagnostic subcomponents were less distinguishable from each other, suggesting that there is little meaning gained from identifying diagnostic subcomponents using the CAELD essay for students below a particular band score. That being said, the raters were able to offer the students advice across all of the proficiency levels, including the lowest scored essays. The CAELD TLU domain needs to be specified even further by identifying a band score threshold for the tasks, in order for the results to provide useful diagnostic feedback about the students’ writing ability. Since the range of essays included in Phase 1 were representative of the EAP course included in Phases 2 and 3, the concerns observed in Phase 1 might be relevant for the other two phases.

Further, the identification of reading diagnostic subcomponents might only be meaningful for students above a certain band score based on possible rebuttal data, which are grounded in rebuttal data from the teacher for warrant C.2, discussed in more detail below. In addition, rebuttal data indicate that reading speed is not measured as a distinct category from the other reading diagnostic subcomponents.
The students’ inability to complete the reading question booklets was labeled as reading speed, but this inability to complete the questions was also used to identify other diagnostic reading subcomponents, such as getting the gist. As such, there would be a correlation between reading speed and any of the other diagnostic subcomponents confounding the meaningfulness of the feedback for teaching and student learning. Diane identified reading speed as a needed skill and targeted it in the course, which the students reported as beneficial for their learning, and thereby meaningful to the CAELD construct.

Indeed, Fox (2009) identified “ability to cope with heavy reading load” (p. 40) in a timed setting based on Banerjee and Wall’s (2006) work, as a CAELD diagnostic subcomponent. For reading speed to be a meaningful subcomponent that is distinct from the others, another task needs to be developed that only measures reading speed.

Table 29

*Meaningfulness Warrants (C.3–C.4) for Assessment Claim C*

<table>
<thead>
<tr>
<th>Warrants</th>
<th>Backing</th>
<th>Rebuttals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Warrant C.3:</strong> The procedures for administering the CAELD enable test takers to perform at their highest level on the assessment (eliciting students’ optimal performance).</td>
<td>• The procedures for administering the CAELD were the same as the CAEL (CAEL, 2008).&lt;br&gt;• The test was administered in the classroom, rather than in the CAEL high-stakes standard testing environment. This situation was less stressful for some students, enabling them to perform at a higher level (Chapter 5).</td>
<td></td>
</tr>
<tr>
<td><strong>Warrant C.4:</strong> The CAELD teacher and students’ learner profiles focus on elements of the EAP course instruction and student learning (utility of the assessment report).</td>
<td>• The diagnostic subcomponents corresponded to skills and strategies taught (Chapter 5).&lt;br&gt;• The diagnostic subcomponents corresponded to skills and strategies that the students focused on when studying (Chapter 6).</td>
<td>• Possible rebuttal: The diagnostic subcomponents only focused on a portion of the curriculum (Chapter 5).</td>
</tr>
</tbody>
</table>

The current CAELD construct underrepresents the targeted skills needed to support teaching and learning in the intermediate-level EAP course. Diane and her students identified vocabulary as an
essential subcomponent, indicating that a separate test or section on vocabulary is necessary. Student
data also suggest the need for a section focusing on grammar. Douglas, who was placed into the
listening targeted group, identified the need for a grammar section. However, Douglas was not aware of
the type of feedback that was received by students whose weakest skill was writing. If Douglas’ weakest
skill had been labeled as writing, he would have received feedback on his grammar, and maybe would
not have seen the need for more focus on grammar. The other skill sections (e.g., reading, listening) of
the CAELD indirectly measure vocabulary and grammar, but if the diagnostic feedback is to be
meaningful for teaching and learning, the construct needs to be expanded to include separate feedback
on vocabulary and, possibly, grammar. As further evidence of construct underrepresentation, the raters’
verbal protocols did not elicit the same number of strengths and weaknesses across the three levels of
essays. This possible construct underrepresentation highlights a distinct difference between tests used
for university admissions purposes and diagnostic purposes within an EAP program. For university
admissions, the construct should be broad enough to estimate the students’ language proficiency
necessary to be successful, but for diagnostic purposes the focus is more specific and related to the
curriculum of the course.

The second meaningfulness warrant (C.2) complements the first by examining the extent that the
CAELD test tasks elicit the appropriate test behaviour for the defined construct. Once again, there was
conflicting evidence supporting and refuting this warrant. Backing data from the three results chapters
maintains that the CAELD essay is an appropriate task for eliciting academic writing, as well as for
providing evidence of students’ strengths and weaknesses. Rebuttal data from Chapters 5 and 6 suggest
that the reading skill section might not be suitable for eliciting students’ strengths and weaknesses.
Interestingly, Diane thought the reading section was too difficult for her students, based on the length
and complexity of the reading text, while students remarked that the reading items did not accurately
assess their reading ability, specifically reading comprehension.
Diane’s comments speak to Messick’s (1989) notion of construct-irrelevant difficulty. According to Messick (1989), construct-irrelevant difficulty describes "aspects of the task that are extraneous to the focal construct make the test irrelevantly more difficult for some individuals or groups" (p. 34). Though reading academic English texts is a part of the construct, the reading section of the CAEL might have been too difficult for some students, thus impeding their ability to answer the questions. These answers are necessary to identify the diagnostic subcomponents, the central piece of the CAELD construct. This potential construct-irrelevant difficulty is a divergence from the CAEL construct, which is to determine a student’s proficiency level. If the students are not able to read the text because it is too difficult, the resulting low scores are meaningful to the CAEL construct.

In contrast to the teacher’s comments, the students’ remarks echo Messick’s (1989) understanding of construct underrepresentation. Specifically, Douglas and Neil questioned whether the reading section actually elicited reading comprehension skills rather than just test-taking skills. The two stakeholder perspectives illustrate the differences between the teacher’s and students’ goals and perspectives. The teacher needs to consider all of the students, with varying proficiency levels and learning needs, in her course. The students’ goals are more individual. Douglas and Neil, who were among the strongest students in the course, only considered the text and the items in relation to their own CAELD experience. It is unlikely that the level of difficulty of the reading text inhibited them from answering all of the questions, since they spoke of the need for more challenging ones. These stakeholder perspectives highlight the need for more research to mitigate these possible threats of construct underrepresentation.

The third meaningfulness warrant articulates that the administration of the CAELD ensures the students are able to perform at their highest ability (warrant C.3; see Table 29). The backing data draw on the fact that the CAELD was administered in the same way as the CAEL, and suggest that the students’ test performance should have been as high as if they took the CAEL. A key difference between the CAEL and CAELD testing contexts is that the CAELD administrator was also the students’ teacher.
Thus the classroom setting for the administration might have been less stressful for students, thereby drawing a higher test performance from them than if conducted in a high-stakes testing context administered by a stranger.

Table 30

*Meaningfulness Warrants (C.5–C.6) for Assessment Claim C*

<table>
<thead>
<tr>
<th>Warrants</th>
<th>Backing</th>
<th>Rebuttals</th>
</tr>
</thead>
</table>
| **Warrant C.5:** The teacher and students interpret the CAELD feedback appropriately (meaningful for stakeholders). | • The students’ perceived strengths and weaknesses matched the CAELD diagnostic subcomponents (Chapter 6).  
• Some students agreed or were content with the CAELD feedback (Chapter 6).  
• The students’ understanding of diagnostic assessment was mediated by course instruction (Chapter 6). | • There were no data collected about the teaching assistants’ interpretations of the learner profiles (Chapter 5).  
• The students did not initially know how to interpret the results to enhance their own studying of English (Chapter 6).  
• Some students did not agree with the results in the learner profile because they did not match their self-assessment (Chapter 6).  
• Students initially had a shallow, score-based interpretation of the results (Chapter 6). |
| **Warrant C.6:** The test developer clearly communicates the definition of the construct, in non-technical language, before the administration of the CAELD. The construct is clearly defined on the learner profile for teachers and students (the construct is communicated to stakeholders). | • The teacher had access to the test specification information and test score users guide for the CAELD (Chapter 5).  
• In administering the CAELD, the teacher communicated the definition of the CAELD, based on the CAEL test administrator’s guideline (Chapter 5).  
• The teacher communicated the construct to the students on the first day—in the course outline and with a presentation about the CAELD—and throughout the diagnostic assessment process (Chapter 5). | • Some students might have been unaware of the importance of trying equally hard in all sections of the test. Students’ comments suggested that they tried harder in particular sections of the test because it was for diagnostic purposes; the subtest scores might have underrepresented their ability (Chapter 6). |

The construct was communicated to the stakeholders through a variety of methods and modes, such as the CAEL website, the introduction on the first day of class, and the administration of the CAELD; it was also communicated before the delivery of the learner profile (warrant C.6; Table 30). The most obvious method of communication is through the learner profile. Based on the backing data,
the learner profile focuses on the aspects of the construct meaningful to the intermediate-level EAP course instruction, and to the students studying, both on their own and in their other university courses (warrant C.4).

The teacher’s and students’ interpretations of the construct varied (warrant C.5; Table 30). Atypical of most testing contexts, the teacher and I (the researcher) took on different aspects of the test developer role, namely, the creation of the learner profiles. In addition, the teacher was the test administrator. Through these multiple roles, the teacher was able to develop a deeper understanding of the test construct and the students’ abilities, which led her to integrate the feedback more or less seamlessly into her instruction. However, the students’ interpretations of the CAELD construct were mixed. The backing and rebuttal data demonstrated that the students’ interpretations of the construct were based on their own self-assessment, previous experiences with the CAEL, and/or the administration of the CAELD. Perhaps most importantly, the students’ interpretations were mediated by the course instruction.

Table 31

*Impartial Warrant (C.7) for Assessment Claim C*

<table>
<thead>
<tr>
<th>Warrants</th>
<th>Backing</th>
<th>Rebuttals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Warrant C.7:</strong> The CAELD does not include response formats or content that may either favour or disfavour some students in the EAP course (impartial format and content).</td>
<td>• Prior research demonstrates that the response formats or content does not favour or disfavour certain test-taker groups (CAEL, 2008).</td>
<td>• Students who are familiar with the CAEL will have an increased familiarity with the response formats (Chapter 6).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Some students were more comfortable with other language proficiency test formats and content (Chapter 6).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• The raters interpreted some of the student essays’ strengths and weaknesses in relation to the perceived effort of the students (Chapter 4).</td>
</tr>
</tbody>
</table>

Despite the fact that the CAELD construct was communicated to the students in a number of ways, rebuttal evidence suggests that the students did not completely understand the need to try equally
hard across all of the skill sections (warrant C.6; see Table 30). These rebuttal data suggest that there was a miscommunication about how the results were going to be used. The students were only informed of how the learner profiles were created shortly before they were given the profiles, but not before they took the CAELD. Therefore, the students did not understand that they were going to be placed into groups (identified in their profiles) based on their overall weakest skill, as identified by the CAELD. If the students were aware of the decision-making process prior to taking the CAELD, they would have been fully informed as to the purpose and consequences of the test and its results.

Table 32

*Impartial Warrants (C.8-C.11) for Assessment Claim C*

<table>
<thead>
<tr>
<th>Warrants</th>
<th>Backing</th>
<th>Rebuttals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Warrant C.8:</strong> The CAELD does not include content that may be offensive to some test takers (culturally appropriate).</td>
<td>• There were no comments from students to suggest that the content was offensive (Chapter 6). • The CAEL has been examined to be culturally sensitive for different test-taker populations (CAEL, 2008).</td>
<td></td>
</tr>
<tr>
<td><strong>Warrant C.9:</strong> The students have equal access to the CAELD content and procedures, and are equally familiar with testing conditions (equal access to assessment information).</td>
<td>• The CAEL website makes information about the CAEL available to test takers (CAEL, 2008). • The teacher introduced the CAELD to the students on the first day (Chapter 5).</td>
<td>• Novice CAEL test takers and students absent from class on the first day would not have learned about the CAELD, unless they went to the CAEL website (Chapter 5).</td>
</tr>
<tr>
<td><strong>Warrant C.10:</strong> The procedures for creating the learner profile are clearly described in terms that are understandable to all test takers (equal access to scoring process).</td>
<td>• The teacher included a presentation on how the learner profiles were created (Chapter 5).</td>
<td>• Students absent from class the day that the student learner profiles were delivered to students would have missed the information about how the profiles were created (Chapter 5).</td>
</tr>
<tr>
<td><strong>Warrant C.11:</strong> Interpretations of the students’ CAELD feedback are equally meaningful for students from different first language backgrounds and academic disciplines (impartial scoring).</td>
<td>• Previous research on the fairness of the CAELD supports this warrant (CAEL, 2008).</td>
<td></td>
</tr>
</tbody>
</table>

175
This discrepancy in timing to communicate the construct and the decisions to students exposes a critical difference between the purpose of the CAEL to the CAELD and the students’ test taking approach. For example, students taking the CAEL for admissions or placement purposes are aware that a high overall band score is needed for their intended goals, and the strategic test taker would likely spend more time on the writing section, knowing it is weighted more. Contrast this example with a student taking the CAELD who wants feedback about a particular skill: His or her energy might be focused on that section, and not the others, with little thought about the overall band score. The second example also demonstrates the multiple interpretations that the term “diagnostic” can have. The student who primarily focused on reading and not writing (Chapter 6) might have expected to receive feedback on all of the skills tested by the CAELD, and not only the weakest by skill band score. As a result, the student only received feedback on writing. Even though there are only six warrants qualifying the meaningfulness category of the interpretations claim, there is rich evidence supporting and refuting the warrants from across the three results chapters.

Impartial. The impartial warrants qualify that the interpretations are consistent across all groups of test takers and are not biased to any one particular group (see Tables 31 and 32). Since there was a difference between the students who had and had not taken the CAEL previously, two groups of novice and experienced CAEL test takers were considered. These two groups likely resulted in uneven interpretations about students’ abilities (warrant C.7). The experienced CAEL test takers were familiar with the administration procedures and test tasks. Thus this more experienced group had an advantage over the novice group, who took the test for the first time in the course, as in or on the CAELD. In a high-stakes setting, such as the CAEL, it is expected that students will have equal access to test preparation materials. However, it is unrealistic and against the spirit of assessment for learning, to expect students to prepare for a test being used formatively (Black & Wiliam, 1998; Rea-Dickens, 2001). This difference in knowledge and experience between the novice and experienced groups creates a potential unfair testing situation for the CAELD. Further, the students’ perceived effort in taking the
test seriously might have biased the raters in their scoring of the CAELD essays. While the same raters who participated in Phase 1 also scored the CAELD essays for Phases 2 and 3, the rationale for this possible rebuttal is based on the raters’ think alouds and interview data and not their actual scoring of the students’ CAELD essays.

Overall, the backing data support warrants C.7–C.11. While the rebuttal data regarding the students missing class and the details about the test are noteworthy from an instructional standpoint, it would only be critical if the CAELD were also used for a high-stakes test, such as certification.

Table 33

*Generalization Warrants (C.12–C.13) for Assessment Claim C*

<table>
<thead>
<tr>
<th>Warrants</th>
<th>Backing</th>
<th>Rebuttals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Warrant C.12:</strong> The observations made are representative of learning tasks in both the EAP course and introductory university courses (generalization to subsequent learning).</td>
<td>- The teacher used a variety of tasks similar to the CAELD in her course instruction (Chapter 5).&lt;br&gt;- Some students spoke of the similarity between the lecture-listening task of the CAELD essay and their tasks in their mainstream courses (Chapter 6).</td>
<td>- Some students identified concerns that the sustained test topic was not closer to their field of study (Chapter 6).</td>
</tr>
<tr>
<td><strong>Warrant C.13:</strong> The criteria and procedures for recording the responses to the CAELD tasks correspond closely to those that were used by the teacher in the EAP course (generalization of scoring to TLU tasks).</td>
<td>- The teacher used a variety of methods to evaluate students’ abilities similar to that of the CAELD (e.g., constructed response) (Chapter 5).</td>
<td></td>
</tr>
</tbody>
</table>

*Generalizations.* The two generalization warrants explore the correspondence between the CAELD test tasks and the TLU domain (Table 33); in other words, they explore the extent to which the CAELD test tasks are similar to the types of tasks that students will experience in the EAP program and in first-year university courses (Bachman & Palmer, 2010; Loevinger, 1957). Overall, the test tasks were similar to those used by the teacher and experienced by the students in their content courses. Rebuttal evidence based on the students’ comments suggests that the test topic was too scientific. The CAELD draws on the use of a variety of academic topics that cover the social sciences and physical sciences,
which is typical for tests that measure academic English ability. However, student dissatisfaction with an integrated test topic is not new, and research demonstrates that the validity of the interpretations of students’ overall performance is not compromised by differences in test topics (see Fox, 2004). I anticipated the generalizations observed between the CAELD test tasks and the TLU domain, since the CAEL construct was modelled on the types of tasks students would be expected to experience in their first-year university courses.

**Relevance.** The relevance warrants qualify how well the test interpretations serve the needs of the decision makers—both the teacher and students (Tables 34 and 35). Data from across the three results chapters converged to provide strong backing regarding the relevance of the interpretations for the teacher’s and students’ decision-making needs (warrants C.14–16). However, rebuttal data from Chapter 4 suggests that the interpretations might be too narrow when translated into advice for students on how to improve their studying (warrant C.14). Along the same line, the CAELD feedback only represented a portion of the curriculum and course goals.

Table 34

**Relevance Warrants (C.14–C.15) for Assessment Claim C**

<table>
<thead>
<tr>
<th>Warrants</th>
<th>Backing</th>
<th>Rebuttals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Warrant C.14:</strong> The students’ identified strengths and weaknesses in writing connect to instructions for how the student can improve (meaningful observations).</td>
<td>• Overall, the strengths and weaknesses observed by raters connected to feedback (Chapter 4).&lt;br&gt;• The types of targeted approaches by students corresponded to the advice offered by the raters (Chapter 4 and Chapter 6).</td>
<td>• The raters provided more feedback about the language category of the scoring guide over the other two categories (Chapter 4).</td>
</tr>
<tr>
<td><strong>Warrant C.15:</strong> The diagnostic feedback maps onto the teacher’s overall course planning (relevant to course instruction).</td>
<td>• Overall, the teacher used the range of diagnostic subcomponents on the learner profile (Chapter 5).&lt;br&gt;• The teacher had a clear idea of what type of diagnostic subcomponent she wanted targeted (Chapter 5).&lt;br&gt;• The course goals were broader than the CAELD construct (Chapter 5).</td>
<td>• The learner profile created included less diagnostic subcomponents than the CAELD (Chapter 5).</td>
</tr>
</tbody>
</table>
The representativeness of the CAELD construct to teaching is an important piece to consider with a learning-oriented test like the CAELD. In a high-stakes, summative testing context, the narrowing of a curriculum would be classified as negative washback (Alderson & Wall, 1993; Cheng, 2008). In a low-stakes, diagnostic setting, the CAELD feedback needs to shape the decisions made by the teacher and students; otherwise the test is not useful (Bachman & Palmer, 1996). However, it would be more disconcerting if the curriculum, teaching and learning, was entirely faithful to feedback that only reports on linguistic skills—a move away from the EAP’s larger goal of supporting students’ academic acculturation into their post-secondary studies (Cheng & Fox, 2008; Evans & Morrison, 2010).

Table 35

*Relevance Warrant (C.16) for Assessment Claim C*

<table>
<thead>
<tr>
<th>Warrants</th>
<th>Backing</th>
<th>Rebuttals</th>
</tr>
</thead>
</table>
| **Warrant C.16:** The diagnostic feedback maps onto the students’ learning goals for the EAP course and university studies (relevant to student learning). | • The writing section of the learner profile was relevant to the students’ decisions on how to improve their essay writing (Chapter 6).  
• Mastery and performance goal-oriented students used the results to direct their studying of academic English (Chapter 6).  
• Some test takers indicated the learner profile was useful for remembering the diagnostic subcomponents they needed for improvement (Chapter 6).  
• The students were able to extend the CAELD feedback to additional strengths and weaknesses in their intermediate-level EAP course (Chapter 6).  
• When mediated by the course instruction, the students found the diagnostic assessment process relevant to their learning needs (Chapter 6). | |

*Sufficiency.* The sufficiency category of the interpretations assessment claim was based on one warrant that elaborates on the stakeholders’ comfort level with using the CAELD feedback for their
decision-making purposes (Table 36). Overall, it seemed that the teacher and students felt comfortable with the sufficiency of the feedback for informing their decisions.

Table 36

**Sufficiency Warrant C.17 for Assessment Claim C**

<table>
<thead>
<tr>
<th>Warrants</th>
<th>Backing</th>
<th>Rebuttals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Warrant C.17:</strong> The teacher and students are comfortable with the diagnostic feedback for their decision-making purposes, when accounting for possible interpretation errors (sufficiency).</td>
<td>• Some of the students based their decisions to target particular aspects of their English learning only on the CAELD feedback (Chapter 6).</td>
<td>• The teacher used other information to form her decisions on how to group students (Chapter 5).</td>
</tr>
</tbody>
</table>

**Decisions.** The third stage of the assessment process, the decisions piece, serves as the bridge between the interpretations about students’ test performance (Claim C) and the consequences (see Claim A). The intended purpose of the CAELD was for the teacher and students to use the feedback diagnostically in order to benefit their teaching and learning. Under the umbrella of diagnostic test use, the teacher and students made many decisions (Tables 37, 38). To situate the decisions claim within the context of diagnostic use, I first summarize the numerous decisions made by the teacher and students.

Even before the course began, the teacher made formal decisions regarding the timing of the CAELD administration, the inclusion of targeted groups to be based on the CAELD feedback, the culminating assignment for the targeted groups, and the use of marks to encourage both participation in the targeted groups and completion of the reflective tasks. Upon scoring the CAELD and receiving the feedback, the teacher created the targeted groups, based broadly by skill band scores (e.g., reading, writing, and listening). She even created a small group (high-level, balanced group) that she personally worked with because the students’ needs were more advanced than the others in the course. Throughout the term, the teacher continued to base some of her instructional decisions on the CAELD feedback, incorporating such tasks as reading speed in the different units in the course. In line with Edelenbos and
Kubanek-German’s (2004) five level classification of diagnostic competence, the teacher exhibited features of the fifth level—“Pädagogischer Takt” (p. 279)—since she was able to “promote optimum student learning, i.e., to further autonomy, language awareness, self-assessment … as an ongoing process” (p. 279). However, there were times when the teacher doubted some of the distinctions between the nuances of assessment and evaluation. As such, the teacher herself was also a learner: She used the opportunity of using the CAELD as a diagnostic tool to enhance her craft.

Table 37

*Decisions Warrant (B.1) for Assessment Claim B*

<table>
<thead>
<tr>
<th>Warrants</th>
<th>Backing</th>
<th>Rebuttals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Warrant B.1:</strong> The CAELD feedback corresponds to the teacher’s and students’ values about teaching and learning (value sensitivity for the teacher and students).</td>
<td>• The teacher incorporated the CAELD into her course instruction, based on principles of AFL and AAL (Chapter 5).</td>
<td>• At various stages in the course, the students saw value in the CAELD feedback for their learning (Chapter 6).</td>
</tr>
</tbody>
</table>

The students’ decisions varied, both by individual and over time. Akin to Black and Wiliam’s (1998) two-staged process of formative feedback, the students in the intermediate-level EAP course first had to perceive a gap between their learning goals (defined by the course or their personal learning goals) and their ability and, second, they had to act on this gap. In this study, the students in the first stage decided whether or not they agreed with, disagreed with, or were accepting of the feedback. Despite the fact that the students seemed to differ in how they responded to the feedback, the students as a group were fairly consistent in considering the feedback throughout the term, whether inside or outside of class. The second stage of the decision-making process resulted in the students employing a number of different approaches to target their weaknesses: reviewing and learning skills, practicing, learning new strategies, and seeking advice from others. By summarizing the decisions made by these two, key decision-makers, I have set the backdrop to appraise the decisions Assessment Claim B.
This assessment Claim B acknowledges that decisions from the results are considered against the value systems held by stakeholders in the education context, and that the decisions made are equitable for all of the students included (Bachman & Palmer, 2010). To examine this decisions bridge between the interpretations and consequences, Assessment Claim B stated, the decisions made from the interpretations of the CAELD (Claim C) reflected the values of the teacher and students and were equitable for all of the students. Four warrants elaborated the decisions assessment claim: two based on value sensitivity and two on equitable decision-making. Below, I discuss these three warrants together because of the connected nature of the evidence across the warrants.

Since the teacher was the primary decision-maker, she was able to control both the instructional use of the CAELD and the alignment between the decisions made and her teaching philosophy—in essence, her value system of teaching, learning, and assessment (warrant B.1). The core values of assessment for and as learning (AFL/AAL) materialized in her day-to-day teaching and overall instruction. Her teaching philosophy was emergent and developing with a wide range of assessment and teaching experiences informing her instructional practice.

In a similar fashion, the students’ previous assessment and course experiences shaped how the student considered his or her own decisions from the CAELD (warrant B.1). For example, David did not understand why he was asked to take a test at the beginning of a course, before he was given the opportunity to learn the material being tested. In this instance, the decision to include the CAELD as a diagnostic assessment at the beginning of the course was counter to David’s values of assessment and learning. Further, the course instruction served as a mediating component in facilitating the students’ understanding of the teacher’s decisions, and perhaps in modifying decisions about their studying practices based on the CAELD.
Table 38

Decisions Warrants (B.2–B.4) for Assessment Claim B

<table>
<thead>
<tr>
<th>Warrants</th>
<th>Backing</th>
<th>Rebuttals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Warrant B.2:</strong> The relative seriousness of classification errors is considered in relation to the existing education values of the teacher and students (consequences of classification errors).</td>
<td>• The teacher sought the students’ reactions to the CAELD feedback and the targeted groups they were placed in (Chapter 5).</td>
<td>• Initially, some students did not see value in the CAELD and the organization of the groups (Chapter 6).</td>
</tr>
<tr>
<td></td>
<td>• The teacher sought additional information to create learning profiles (Chapter 5).</td>
<td></td>
</tr>
<tr>
<td><strong>Warrant B.3:</strong> The teacher’s primary decisions to group the students are made using decision rules, and no other considerations are used (equitable decision rules).</td>
<td>• Possible: to create the learner profiles, the teacher sought additional information about the students’ abilities (Chapter 5).</td>
<td>• Possible: to create the learner profiles, the teacher sought additional information about the students’ abilities (Chapter 5).</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Warrant B.4:</strong> The students are informed about the procedures for how the decisions are made, and the decision-making process occurs in this way (stakeholders informed about decision-making process).</td>
<td>• The teacher gave a presentation to the students about how the learning profiles were created (Chapter 5).</td>
<td></td>
</tr>
</tbody>
</table>

It is unusual to consider the second warrant (B.2) regarding classification errors for a test used within a classroom setting. Since the placement of students into targeted groups was based on their CAELD test performance, classification errors were possible. Such errors are typically addressed at the trialing stage of the test development process (Bachman & Palmer, 2010). However, the teacher’s decision for placing students into the groups was emergent and based on the range of skills identified within each course section in combination with the students’ overall needs, and therefore it was impossible to examine the possibility of errors a priori. The severity of the classification errors should have been relatively low considering all students were given the opportunity to learn from the very fact of being placed in a group. To counter the possibility of classification errors, the teacher met with the students and ensured that they were satisfied with the placement decisions. The classification warrant is not necessarily a good fit for the current study, since the decisions were largely based on the students’
weakest skill, and not on a cut score. For example, it is likely that there were students in the listening group who had different listening band scores, but were in the group because it was their lowest skill score. That being said, this warrant expands on how the teacher used the CAELD when making major decisions regarding placing students into groups.

Unique to the decisions assessment claim, the evidence serving as backing for warrant B.2 (value sensitivity), seen above, also forms possible backing and/or rebuttals for warrant B.3 (equitable decision-making) depending on the perspective taken of test fairness. The possibility that the same evidence might be used for the opposing ends of the same warrant underscores the tension inherent when using a large-scale assessment in a classroom setting for diagnostic (formative) purposes. The teacher sought additional information to inform her decisions about how to place students in the groups (possible backing). The process by which she drew on additional information was not systematic, and therefore could be seen as rebuttal data. At issue is what Kane (2010) refers to as substantive fairness, which he defines as the requirement that “the score interpretation and any test-based decision rule be reasonable and appropriate, and in particular, that they be equally appropriate for all test takers (at least roughly)” (pp. 178-179). His definition and surrounding discussion makes clear the subjective nature of determining what is appropriate and fair for all test takers. From a large-scale assessment perspective, the idea of fairness brings forth the notion of consistency in decision-making across large test taker populations, while a classroom-based assessment perspective might consider the individual and what is best for his/her learning needs. Bachman and Palmer (2010) acknowledge there does not need to be consistency between the decisions made to the value systems of the educational context and the stakeholders; they do, however, argue that the contradictory perspectives should at least be considered. The challenge here is to decide what perspective should be given precedence.

The decisions piece of the AUA exemplifies the bridge between the interpretations about students’ abilities and the consequences, marking the pivotal point when the test users act on the
CAELD feedback, resulting, it is hoped, in beneficial consequences for decision-maker and for other stakeholders (see Claim A).

Table 39

**Beneficial Consequences Warrants (A.1-A.3) for Assessment Claim A**

<table>
<thead>
<tr>
<th>Warrants</th>
<th>Backing</th>
<th>Rebuttals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Warrant A.1:</strong> The CAELD is efficient to administer and score within the course schedule (beneficence of administration).</td>
<td>Some students viewed the CAELD as an opportunity to learn and approached the administration accordingly (Chapter 6).</td>
<td>The teacher found the CAELD time-consuming to administer and score (Chapter 5).</td>
</tr>
<tr>
<td><strong>Warrant A.2:</strong> The students have no adverse experiences with taking the CAELD (beneficence of administration).</td>
<td>Despite being low-stakes, some students approached the administration of the CAELD as they would a high-stakes test (Chapter 6). The CAELD took approximately three hours for the students to complete (Chapter 5).</td>
<td></td>
</tr>
<tr>
<td><strong>Warrant A.3:</strong> The CAELD feedback informs the teacher’s instruction of individual students, the whole class, and also informs her teaching, overall (beneficence for instruction).</td>
<td>When the teacher reflected on her experience of creating her own diagnostic assessment, she noted that being able to use a pre-existing test was more convenient than having to create one (Chapter 5). The teacher modified her instruction (separate from the targeted groups) according to the CAELD feedback (Chapter 5). The teacher was able to group the students according to the CAELD feedback (Chapter 5). The teacher’s use of the CAELD exhibited features of formative assessment (Chapter 5). The teacher used the CAELD feedback to inform her instruction of future EAP courses (Chapter 5).</td>
<td>Possible: The teacher relied on instructional support to help facilitate her use of the CAELD (Chapter 5). Possible: The teacher remarked that even with support, she was unable to give the students the individual assistance they truly needed (Chapter 5). Despite all of the work on the part of the teacher, she was not the primary user of the feedback (Chapter 5).</td>
</tr>
</tbody>
</table>

**Beneficial consequences.** The final stage of the assessment process considers the outcomes from the decisions made by the test users—in other words, the intended and unintended consequences of the test. The intended purpose of the CAELD is that it benefits the teacher’s instruction and the students’
studying and learning of academic English. The beneficial consequences assessment claim is grounded in Kunnan’s (2004) notion of beneficence—that use of a test should be positive for society and should, most importantly, do no harm. Therefore, to articulate the intended purpose of the CAELD, Assessment Claim A stated, “The consequences of using the CAELD for diagnostic purposes in an EAP classroom will be beneficial to the teacher and students.” Five warrants elaborated the claim (see Tables 39, 40, and 41).

The administration of the CAELD was an integral part of the teaching and learning in this study. Thus I discuss the corresponding warrants (e.g., administration and teaching/learning) together for the teacher, and then for the students. To conceptualize my rationale for classifying the evidence as backing or rebuttals for the teacher and her instruction, I draw on Jamieson’s (2011) notion of classroom assessment as interwoven into the fabric of teaching. Framed in this way, the use of the CAELD for diagnostic (formative) purposes is considered beneficial for the teacher if she can seamlessly integrate the assessment into her practice (Shepard, 2006), but not if integrating the assessment is counter and jarring to her typical approach (Davison & Leung, 2009). Further, the use of the CAELD is beneficial if it strengthens course instruction, but is detrimental if it diminishes the effectiveness of instruction for students (Moss, 2003).

The CAELD was beneficial to the teacher, personally, and to her overall instruction (warrant A.3). The teacher spoke about her overall experience as positive. She was able to use the CAELD feedback in a number of ways, from organizing students into targeted groups to identifying diagnostic subcomponents, which she included both when working with individual students and with targeted groups. The teacher was also able to draw on her experience to reflect on her future teaching. The administration of the CAELD at the beginning of the course was probably one of the greatest causes of tension for the teacher (warrant A.1). Although the warrant has only one rebuttal, it stems from extensive qualitative data in Chapter 5, which clearly depicts the teacher’s frustration with the time-consuming nature of administering the test and creating the learner profiles. It was the teacher’s decision
to include the administration of the CAELD during class time because she wanted the students to value the CAELD and see it as an integral part of the course. However, by choosing class time for the administration and the targeted groups, the teacher gave up valuable teaching time, which she later remarked was something that she had not anticipated. The teacher felt this was a loss of teaching time since the administration of the CAELD took one whole class period, three hours, which in turn meant she had to remove a lesson from the overall course plan. This loss of teaching time could be seen as disadvantageous for students. Another challenge for the teacher was that, while she put time and effort into creating the learner profiles, she was not their primary user, but rather the teaching assistants, who were responsible for leading the targeted groups.

I characterize student learning as a dynamic and co-constructed process among the student, the teacher, and the other students. The study was situated within a particular classroom context, with learning goals set by the teacher. One of the learning goals set by the teacher for her students, which is also suitable for evaluating the CAELD, was for students to become independent and autonomous learners. When taking a broad perspective of learning, there was considerable positive evidence to suggest that the use of the CAELD yielded benefits for the students’ studying of academic English including reported increased self-awareness, identification of learning goals, learning new studying habits, to practicing a skill (warrant A.4). Beyond the individual nature of the students’ studying, many of the students reported the positive benefits from interacting with their peers in the targeted groups. From a socio-constructivist perspective (Lantolf, 2005), this peer group interaction highlights the co-constructed nature of learning that took place for the students. For instance, both Elaine and Don commented positively on being able to interact with their peers and receive feedback about their writing from one another. Students with different learning styles seemed to benefit from the decisions based from their own use of the CAELD feedback. Students who were already quite autonomous learners, such as Neil, perhaps benefited more from taking the CAELD than students who were less independent, such as Kevin.
Table 40

**Beneficial Consequences Warrant (A.4) for Assessment Claim A**

<table>
<thead>
<tr>
<th>Warrants</th>
<th>Backing</th>
<th>Rebuttals</th>
</tr>
</thead>
</table>
| **Warrant A.4:** The students use the diagnostic feedback to modify their studying of academic English (beneficence for student learning). | • The consideration of the CAELD feedback remained constant throughout the EAP course (Chapter 6).  
• The students incorporated the feedback into their studying of academic English in a variety of ways and at different times (Chapter 6).  
• Based on the CAELD feedback, some of the students adjusted their habits for studying academic English (Chapter 6).  
• Some of the students identified learning goals for themselves that were based on the CAELD feedback (Chapter 6).  
• The students commented that—due, primarily, to the teaching—their experiences with the diagnostic assessment process were mostly positive (Chapter 6). | • No data were collected about the students’ test performance at the end of the course.  
• Possible: Some of the students commented that they would not have practiced their weaknesses if it were not for the targeted groups (Chapter 6).  
• Possible: The beneficial use of the CAELD feedback by students was mitigated by the teacher’s diagnostic competence (Chapter 5).  
• Possible: If the CAELD feedback were inaccurate, the students might have been encouraged to study skills they did not need to study (Chapter 8). |

The CAELD might or might not have led to the students’ improved metacognition and self-regulation skills. However, there were no data collected that could support such assertions. Further, it is unknown if the students’ targeted studying resulted in an improvement in their proficiency, either overall or for particular skills. The students might or might not have continued to use the strategies they learned during their targeted practice as identified by the CAELD feedback once the course finished.

The students’ experience with the administration was mixed (warrant A. 2). Some students were clearly overwhelmed by the testing-like context, the length, and the verbal instructions—conditions that were similar to those of any large-scale, standardized test. Doe and Fox (2011) noted that students’ strategy use and approach to the CAEL varied across testing contexts, such as the classroom; a practice test; and a high-stakes, live testing situation, depending on the stakes. More research is needed to investigate
whether or not students experience the same level of anxiety with a teacher-developed assessment, administered at the beginning of the course for diagnostic purposes.

Thus it would seem that the CAELD was beneficial for both teaching and learning. However, the possible rebuttal data for warrants A.3 and A.4 question the utility of the CAELD in the absence of the additional support provided to the teacher, or when teachers do not have the same level of diagnostic competence. Without the teaching assistants, the teacher would not have been able to set up the targeted groups. Even as the researcher, I was able to provide some administrative support for the teacher by organizing files. The teacher’s competence with assessment came through in her interpretation of the students’ test performance, as well as in her ability to conceptualize the results for appropriate use. Moreover, the students’ reported positive experiences of the diagnostic assessment process were attributed to the teaching, and not the CAELD.

Table 41

<table>
<thead>
<tr>
<th>Warrants</th>
<th>Backing</th>
<th>Rebuttals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Warrant A.5</strong>: The teacher and students experience no adverse effects from the decisions made from the diagnostic feedback (beneficence of decisions).</td>
<td>• Overall, the majority of the students seemed to agree with the decisions made based on the CAELD feedback (Chapter 6).</td>
<td>• Overall, the students’ comments about the diagnostic assessment process were positive (Chapter 6).</td>
</tr>
<tr>
<td></td>
<td>• One student, Dennis, indicated that he wanted to work on his reading and might have felt that being placed in the writing group was a waste of time (Chapter 6).</td>
<td>• One student, Dennis, indicated that he wanted to work on his reading and might have felt that being placed in the writing group was a waste of time (Chapter 6).</td>
</tr>
</tbody>
</table>

| | The CAELD was included as a completion grade for students and therefore had stakes attached to it (Chapter 6). |

The majority of the students did not seem to have any real adverse effects on their learning resulting from the decisions made by the teacher. That being said, rebuttal data from Chapter 6 highlight the importance of considering the students’ learning goals based on formative assessments, such as the CAELD. Although Dennis knew he was weaker in writing, he wanted to focus on his reading ability, his strongest skill according to the CAELD. He perceived reading as more important to his future studies in engineering. The decision to place a student in a group counter to his desires raises questions regarding
the goal of EAP instruction, the purpose of diagnostic assessment, and the need to consider students’ long-term learning goals. When framed from a beneficial consequences perspective, the instruction Dennis received in the small group most likely enhanced his writing skills. Nevertheless, Dennis might have tried harder in the reading group.

**Examining the CAELD**

The purpose of this study was to examine the appropriateness of the Canadian Academic English Language Diagnostic (CAELD) assessment for diagnostic purposes in an English for Academic Purposes (EAP) program. With the study’s purpose in mind, I am cautious to support the use of the CAELD for diagnostic purposes in the EAP program within the context of this study. While there are notable strengths of the CAELD that lend itself as a diagnostic tool, I am acutely drawn to the concerns of construct representation and the overall feasibility of the test as a teaching tool. I recommend a few changes for the test developer to consider that would enhance the validity and reliability of the test for this formative purpose. In addition, I recommend further research to investigate the validity of the test for diagnostic purposes, which could ultimately lead to significant changes to the test structure.

The main support for the CAELD is in its use. The backing for Assessment Claims A and B demonstrates that the decisions and consequences from the CAELD benefited teaching and learning, with the success largely attributable to the teacher and the teaching supportive context. The teacher drew on the learner profiles to group the students into targeted groups based on their profiles and inform her future practice. Within these targeted groups, the students completed tasks focused on their weaknesses. Backing data provided based on the teacher and student data suggest that the overall experience was positive and beneficial.

Additional support for the CAELD as a diagnostic tool is seen in Assessment Claim C. The diagnostic subcomponents mapped onto the teachers’ teaching suggesting that the CAELD was relevant to the intermediate-level EAP course. The students’ perceived strengths and weaknesses corresponded
with the diagnostic subcomponents, lending support of content validity. It is logical to identify support for CAELD in an EAP program, since the test as the CAEL is a well-recognized test of EAP.

My concerns about the CAELD used for diagnostic purposes are rooted in construct representation. While the content of the CAELD is relevant to the intermediate-level EAP course, rebuttal data suggested not all of the subcomponents were appropriate; conversely, the test only measured a portion of the instructional and learning needs of the teacher and students. For instance, speaking was not included in the test. These concerns about construct representation combined with the length of the test need to be heavily weighted against the positive benefits. The administration of the CAELD took a full class, in addition to the make-up class when the students missed the test. This is time taken away from the students’ opportunity to learn. If the CAELD were to continue with a diagnostic purpose, the test developer might want to consider an online approach to administration, which may, in turn, facilitate scoring the CAELD into a one-step process, and reduce the reliability concerns seen in Assessment Claim D. The length of time involved in scoring, evaluating the students’ test performance, and creating the learning profiles, is onerous or burdensome. Such an intensive process requires considerable commitment and knowledge from the teacher.

Thus, for the CAELD to continue as a diagnostic tool, the test developer needs to examine how to make the CAEL accessible and feasible for classroom use. In addition, the construct definition of the subcomponents needs to be expanded and refined to better suit the stakeholders’ needs. To this end, I recommend that further research investigate the use of the CAELD across larger test-taker populations or by different teachers and student populations. Based on additional research, the test developer may want to consider modifying the test structure to allow for more flexibility in including diagnostic subcomponents based on the test users’ needs, the teacher and students.

Embedded in the evaluation of the CAELD is the setting—the stakeholders, the proximity of the CAEL, and myself, the researcher. All of these factors contributed to the use of the CAELD and the interpretation. I caution the readers that this study presents an incomplete or partial validity argument.
Drawing on three stakeholder perspectives provided tremendous depth to the assessment claims.

Through transparency in my description of the methods used (Chapter 3) and findings (Chapters 4-6), combined with the soundness of the assessment claims supporting the AUA (this chapter), the test developer and test users will be able to arrive at their own conclusions about the use of the CAELD.
Chapter 8: Conclusion and Implications

The act of writing a dissertation is deeply personal. Although this writing is primarily written as an objective third-person narrative, my beliefs and interests are at the heart of it. Nash and Bradley (2012) emphasize the important role a writer’s beliefs play in scholarship. I began my graduate career with questions about teaching and learning, and later added assessment to this relationship as the link between teachers and students. Through this dissertation I have observed the dynamic and multifaceted ways in which assessment connects teachers and students. The Canadian Academic English Language Diagnostic (CAELD) served not only as a communication tool between the teacher and students, but also as an aid for teacher and students to achieve their individual goals. Beyond the interplay among teaching, learning, and assessment, I observed the importance that context plays in interpretation and assessment. In the study, that context was grounded in the values held by key stakeholders—raters, teacher, and students.

These multiple stakeholder perspectives provided converging and conflicting accounts as backing and rebuttals for the evaluation of the Assessment Use Argument (AUA) for the CAELD, seen in Chapter 8. In this final chapter, I delineate the findings from the three results chapters and the AUA as implications for testing, as well as for teaching and learning. At the heart of each of these issues is the consideration for test use. This study is one of only a few to examine the consequential use of diagnostic assessments in the language classroom (e.g., Jang, 2005), and the only one I have encountered that examines a test retrofitted for diagnostic purposes, or draws on the most recent version of the AUA (Bachman & Palmer, 2010) as a guiding framework for exploring beneficial consequences in-depth. These unique features offer insight into the beneficial use of tests in learning contexts, and the merits of using argument-based approaches to test validation. Throughout my discussion of these implications, I acknowledge the limitations of the study and point out potential areas for future research.
Beneficial Consequences in situ

An integral component of the AUA included investigating the beneficial consequences of using the CAELD for diagnostic purposes. In the literature, there is considerable debate about whether consequences stemming from test use should be included in the validation process or even conceptualized as part of validity (e.g., Kane, 2010; Lissitz & Samuelson, 2007; Messick, 1989, 1996). This study demonstrates that consequential validity is an integral piece to the validation process of a test with a learning-oriented test use. The beneficial (and negative) consequences of test use are context dependent, because the enabling factors are embedded in a learning setting. The CAELD test users were a teacher and students who were affected by the consequences of decisions made based on the test results. This duality of the test users’ role—both as decision makers interpreting the results and as recipients of the consequences—demonstrates the fluidity of the assessment process and the necessity to consider consequential validity alongside the interpretations and decisions regarding students’ test performance. Due to this desire to positively influence teaching and learning through the use of a test, as in the case of the CAELD, there are a number of similarities, as well as fundamental differences, between this study and washback research.

If the aim of a learning-orientated assessment is to improve teaching and learning, it is essential to investigate whether or not such consequences resulted from the test. However, test developers are reluctant to examine the consequential validity of test use, since they cannot control the actual use by test users (Kane, 2010). Test developers’ reputations are at stake, and their reluctance may stem from concern over the repercussions of negative, unintended consequences: Since test developers are corporations, they are vulnerable to censure and possible legal action (McNamara & Roever, 2006). Though test developers are unable to control test use, they can inform test users, to the best of their ability, of the consequences that may result from the possible decisions in order to facilitate the appropriate use of the test (Bachman & Palmer, 2010). In addition, test developers, researchers, and other stakeholders should be open to the multiplicity of possible uses for test results in a learning setting.
The CAELD was used for diagnostic purposes. Nevertheless, the actual day-to-day use varied greatly by the teacher and across the students, largely based on a combination of their previous assessment experiences, their interpretation of the feedback, and the support available to the teacher and students. A key element of the day-to-day use that spoke to the beneficial consequences was the teacher’s ability to seamlessly incorporate the CAELD feedback into her course. Rea-Dickens (2006) identifies two assessment roles that teachers adopt: (a) identifying students’ proficiency levels, and (b) being sensitive to the students as learners and to their emerging needs. The range of decisions employed by the teacher suggests that the CAELD feedback supported the teacher in both these roles. When examining the effects of test use, the researcher must be open to the range of ways that test use may be beneficial for a teacher. If the CAELD feedback was used by another teacher under similar conditions, the actual use could nonetheless vary, due to variations in teaching styles and levels of experience with assessment. Thus a limitation of this study was that it only looked at one teacher’s use, which may not be representative of other teachers’ methods. However, my in-depth examination of one teacher’s use afforded me a deep understanding of how diagnostic feedback can be incorporated into classroom instruction.

A teacher’s level of assessment literacy is critical to the beneficial use of an externally developed test for formative purposes. Formative instructional decisions should be based on multiple sources of data, which can include large-scale test results in combination with teacher-developed assessments. If teachers do not understand the construct of the large-scale test, they will not be able to adequately integrate the results into their daily instruction alongside their own teacher-developed assessments. In this study, the course instructor questioned the CAELD feedback. I feel this questioning was necessary, as it allowed her to consider the CAELD feedback in tandem with her own assessments of the students’ abilities and thus arrive at decisions about which she felt confident.

The students’ use of the feedback was dependent on their perceptions of and previous experiences with the CAEL, as well as on their own orientation to diagnostic assessment. It was
essential that data be collected from multiple sources over the entire course, in order to observe if the course instruction played a mediating role in the students’ interpretation of the CAELD feedback. The students referred to the connections made between practice tasks, strategies to overcome their weaknesses (e.g., listening during a lecture), and additional resources sought for improving their English. Interestingly, the raters provided similar advice to the students, highlighting the importance of raters who have knowledge about teaching (see Knoch, 2009), and who can help convert test results into meaningful feedback for students, thus providing them with the next steps for their learning. These findings suggest that diagnostic feedback needs to be accompanied by an instructional component in order to facilitate students’ interpretation and use of it. Other researchers have made similar observations. For example, to increase students’ uptake of the Diagnostic English Language Needs Assessment (DELNA) program, staff will meet with students in-person to explain their results to them (Read, 2008). Another consideration is the differing perceptions students have about diagnostic assessment. At the beginning of the course, the majority of the students interviewed had a more summative or evaluative perspective of diagnostic assessment—even though it was defined for them in class with a formative orientation. If students are expected to be active participants in assessment for and as learning settings, then they should be equipped with a diagnostic competence similar to the competence that Edelenbos and Kubanek-German (2004) describe for teachers. Such a competence will provide them with a framework for interpreting the feedback. This assessment competence for students could be achieved through course instruction, as was done in this study, or through accompanying materials that provide the necessary resources and communication links for students to make meaningful connections between the feedback and their learning.

Test developers should become more involved in facilitating additional support for the teachers and students, in order to encourage beneficial consequences from the test use. The teacher in this study received support from the teaching assistants available to her: They helped her to develop tasks that helped the students practice and target their weaknesses. Similarly, by being exposed to the tasks, the
students were able to make the connections between their feedback and activities designed to help them improve. If the test developer is able to provide accompanying materials that provide targeted practice tasks for students, this would hopefully result in greater uptake and use of the feedback than if additional support was not provided. Moss (2003) advocates that assessments used in a classroom setting must have beneficial consequences. Test developers should be working towards developing and promoting tests for beneficial use in classroom settings by teachers and students to the best of their abilities.

This study shares a number of similarities and differences with washback research (e.g., Cheng, 2005; Shohamy, Donitsa-Schmidt, & Ferman, 1996; Wall, 2005). Defined simply, washback refers to the “effects of tests on teaching and learning” (Wall, 1997, p. 291). There are two key features of washback research that relate to large-scale assessments being used for formative purposes: (a) the paramount importance of teaching (and learning) to the process, and (b) the individual and context-dependent nature of examining test use (Cheng, 2008). If the majority of the students had indicated that they felt the CAELD (or the learner profile with the CAELD feedback) was the most important factor in the diagnostic assessment process, this would have been a cause for concern. The test should not be the most powerful influence on teaching and learning; teaching is what ultimately transitioned the CAELD from being a test—a stand-alone instrument—to being an assessment process that facilitated improvements in instruction and students’ studying. Further, since a test like the CAELD is also used for high-stakes decision-making in the same context, care should be taken to ensure that the students’ perceptions and anxiety do not transfer to the use of a test for a lower-stakes purpose. Finally, as has been seen in numerous washback studies (e.g., Cheng, 2005; Qi, 2005), the CAELD feedback did influence the content of the teachers’ instruction.

However, it is unknown if the CAELD also influenced the teacher to modify her teaching style. The washback literature would suggest that it likely did not, but more research is needed to confirm this assumption. A fundamental difference between this study and washback studies is the positioning of the assessment in relation to teaching and learning. In most washback studies, high-stakes testing occurs at
the end of a course to influence teaching and learning. Hence, the name washback originally derived from backwash—the effect that receding waves have in creating a powerful current (Barber, 2004). Though there are similar goals of propelling learning and teaching, the direction of influence and the stakes are different. There were no stakes attached to the CAELD, other than those decided by the teacher. My being in the classroom might have produced a research effect in that it might have influenced the teacher to use the feedback more than if I were not present. However, I do not think this is the case, since the teacher intended to continue using diagnostic assessment in her teaching even though she had little access to additional support. I was able to observe how the course instruction adjusted the students’ interpretation of the feedback as the course progressed, and observed that it was the situation and stakeholders, and not the test itself, that brought about the beneficial consequences. Much as it is the actors who bring a script to life, it is the teachers and students who bring about the appropriate use of an assessment.

Validity Argument(s)

Establishing a validity argument for test use is not new. Since Cronbach and Meehl (1951) first proposed the idea of construct validity, researchers have made attempts to establish inferential links between their research and findings to the often abstract judgments made about students’ test performance. In contrast, argument-based approaches to validation offer an explicit framework for connecting empirical data to such inferences; in this study, I adopted the AUA. While the framework has many strengths, there are also some areas that need improvement. Based on the findings, I propose three new warrants be included to expand the utility of the AUA for examining the beneficial consequences of a large-scale assessment used for formative purposes. Finally, I recognize the value and insight that other frameworks might bring to validating the CAELD.

Of the various argument-based approaches to validation in general education and language testing (e.g., Chapelle et al., 2007; Kane, 2006; Mislevy, Steinberg, & Almond, 2003), the AUA is the most appropriate for the aim of this study because of the weight and importance given to beneficial
consequences. In Claim A, I was able to outline the numerous consequences for the teachers and students that arose from the use of the CAELD in the intermediate-level EAP course (see Tables 39–41). This argument-based approach was appropriate for mapping out the qualitative data regarding stakeholders’ use and perceptions onto the four stages of the assessment process. Stakeholders’ perceptions are typically additive to quantitative internal validity analyses (e.g., equating, scaling, item analyses), rather than central to test validation (DeLuca, Cheng, Fox, Doe, & Li, submitted). However, the findings from this study demonstrate the value that stakeholder data can offer to the interpretation stage (Assessment Claim C; see Tables 28–36) as well as to the consequential outcomes.

Another strength was that the iterative and fluid nature of test use and teaching and learning was reflected in the AUA. McNamara and Roever (2006) argued that one of the greatest challenges facing validity researchers is the clear separation between the psychometric, positivist orientation to interpretation of test taker performance and the messy, socially constructed nature of consequences. In reality, teaching, learning, and assessment use are dynamic and situated within a particular culture of learning. One of the most important consequences for the students in the study was the instruction they received based on the CAELD feedback, which in turn influenced their re-interpretation of the feedback. Without including beneficial consequences alongside the interpretation piece in the validation process, as was done with the AUA, we would not have gained this insight. A major threat to validity is that score interpretation is sometimes narrowly perceived as occurring only before test use and subsequent consequences (e.g., Lissitz & Samuelson, 2007), thus limiting the wealth of information gained from seeing score (and feedback) interpretation as an iterative process.

The third strength was that being able to organize the data into bullet points provided transparency about the strength of the argument, and allowed the test developers and test users to arrive at appropriate judgments about the CAELD. A drawback in using the AUA was the oversimplification of data and redundancy of the warrants. While I particularly appreciated the pragmatic value that the AUA provided when mapping out the data to support the claims, its categories were redundant, in that
the same evidence was used for different warrants, both within a single assessment claim and across
claims; redundancy was also evident in the meaningfulness warrant category, where there were areas of
considerable overlap across the warrants.

There are three warrants that I think should be added to the AUA for future evaluations of large-
scale assessments for learning-oriented purposes. The first considers the test developer’s perspective. A
limitation in this study was that data were not collected about the test developer’s perceptions regarding
the use of the CAELD. As was discussed previously, the test developer is an important stakeholder with
a vested interest in the appropriate use of the CAELD. It is the test developer’s reputation that is
attached to the test use. As an extension of this warrant, I believe that if a large-scale assessment is to be
used formatively, the test developer also needs to provide materials that facilitate and support teaching
and learning beyond the feedback. The students were able to beneficially use the CAELD feedback in
their targeted groups, where they practiced their weaknesses through supplemental tasks. Test
developers could use their knowledge to develop supplemental tasks for students, like the ones used in
the targeted groups. My hesitation in such an approach is that, rather than being used as a springboard
for expanding learning, these materials could be used to teach to the test, thus narrowing learning. I
would also recommend that there be explicit communication about the appropriate uses of such
materials. Therefore, the warrant should articulate the test developers’ perspective of the intended use of
both the test and the supporting materials provided by the developer to bring about that use.

The second warrant I recommend for the AUA is focused on the decisions made using the test
results. The current structure of the decisions assessment claim (in this study, Claim B; see Tables 37–
38) assumes that the types of decisions are determined beforehand, as is needed in a high-stakes testing
context. Prior to the use of the CAELD, the teacher had made a number of decisions about how she was
going to use the CAELD; however, she made many others while scoring the CAELD, and more still
once all of the profiles were compiled. In other words, there were numerous decisions made throughout
the entire process, from before the CAELD was administered until the course finished. Further, all of the
decisions made by the students were iterative. This warrant would establish whether or not of the test users’ decisions are congruent with the overall formative purpose of a test.

The third warrant I propose relates to the teachers’ and students’ knowledge and experiences with assessment. The findings from this study, along with other research (e.g., Volante & Beckett, 2011; Xu & Liu, 2009), demonstrate the importance that assessment experience and literacy has in facilitating and restricting appropriate use of feedback. Thus a warrant is needed that qualifies the range of stakeholders’ experiences with assessment alongside their knowledge of it. Such information would help identify whether or not more training is needed to support stakeholders’ interpretation and use of the test results. Further, emphasizing teachers’ assessment knowledge will give them agency and voice to bring about positive change regarding the use of tests in their classrooms (Boud, 2007; Rea-Dickens, 2004; Taylor, 2009).

Given the dynamic nature of assessment use in a learning setting, I recommend that alternative approaches to validation also be adopted when evaluating such test use. There have been a few studies that draw on narrative and hermeneutic approaches (e.g., Jang, 2005; DeLuca, 2010, 2011) to approach validation research when using an argument-based structure. Research that is sensitive to the dynamic and socially constructed nature of learning will help realize the meaning created by stakeholders. I believe that, when used appropriately, language tests can be an invaluable resource for learning. We are only beginning to realize their potential; one way to gain further understanding is to evaluate their utility against established validity frameworks.

**A Way Forward**

This study is not without limitations. First, I drew on a single case to examine the use of an assessment retrofitted for diagnostic purposes within one EAP program. If the study were to take place in another context, I expect there would be different findings due to the unique situational and program features that influence test use. Second, drawing on the AUA greatly influenced the interpretation of the findings and the evaluation of the CAELD. As stated previously, I recommend that future research
examine the use of a large-scale assessment for diagnostic purposes from a variety of validation frameworks. Third, this study took place within an EAP program context with international students whose language backgrounds were predominantly Arabic and Chinese (including Mandarin and Cantonese). More research is needed to explore how diagnostic assessments can be used to inform teaching of diverse L2 students with a range of language backgrounds and with experience in an English-medium, K–12 educational system prior to their university studies.

Despite these limitations, there are implications for future research and practice related to different aspects of test purpose and use. At a theoretical level, it is important to define the test purpose and the test use. While there is considerable overlap between the two concepts, there are also crucial differences. The purpose of a test relates to how the test performance is distilled and represented in the score reports, or in the case of this study, the learner profiles; test use, however, considers how those reports are interpreted and used by the test users. The nuances of test purpose and test use will be realized through multiple studies across a variety of contexts. Thus more research is needed that examines alternative ways that stakeholders use diagnostic assessments. It is hoped that there will be variation across contexts, since the process would be expected to match a teacher’s personal style.

At a practical level, the multiple roles that stakeholders can occupy within a testing and teaching context should factor into the retrofit of an assessment for a learning-oriented purpose. When an assessment is to be used for another purpose and is already used within the same context for higher-stakes purposes, good communication is key: Stakeholders must be made aware of the differences between the two purposes, albeit for the same test. In addition, more research is needed to investigate and understand how teachers and learners use diagnostic information, and how to better support both.

Messick (1989) described consequential validity as messy. He was right. Yet it is within this messiness that the magic of teaching and learning happens. Future language testing research will continue investigating how this dynamic interaction occurs across educational contexts, making this an exciting time to be a researcher in this area.


Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635–694. doi: 10.2466/pr0.1957.3.3.635


Appendices

Appendix A: CAEL Score Report

The CAEL Assessment Score Report provides a profile of English proficiency in the four skills of academic listening, reading, writing and speaking.

CAEL Assessment results range from band level 10 to band level 90. Each band score corresponds to a descriptive statement summarizing the level of English of a test taker.

**CAEL Band Descriptors**

80-90  Expert User: Demonstrates exceptional competency required for academic English use. Is fluent, accurate, flexible, and adaptable in the academic setting.

70   Adept User: Demonstrates high level of competency required for academic English use. Is fluent, accurate, flexible, and adaptable in the academic setting.

60  Competent User: Demonstrates satisfactory competency in using academic English. Minor limitations in fluency, accuracy, and flexibility in the academic setting.

50  Competent but Limited User: Demonstrates a degree of control in using academic English but fluency, accuracy, and flexibility are somewhat limited in the academic setting.

40  Marginally Competent User: Demonstrates uneven control in using academic English. Fluency, accuracy, and flexibility are impediments to overall competence in the academic setting.

30  Limited User: Demonstrates constrained competency in academic English use. Noticeable problems in fluency and accuracy, and not sufficiently flexible in the academic setting.

10-20 Very Limited User: Demonstrates severely constrained competency in academic English use. Insufficient fluency, accuracy, and flexibility in the academic setting.

**How are the band scores interpreted?**

10-40 need to increase your level of academic English before you meet admission requirements for Canadian University degree programs.

50 may meet academic English language requirements for admission to a few Canadian University degree programs.

60 meet academic English language requirements for admission to some Canadian University degree programs.

70-90 meet academic English language requirements for admission to Canadian University degree programs.

CAEL Assessment test results are maintained in strictest confidentiality. Official score reports are released to other institutions only by written request of the test taker. To ensure the authenticity of CAEL Assessment score reports, institutions are urged to require official copies to be mailed directly from the CAEL Assessment Testing Office.
Appendix B: Summary of Skill Subcomponents Augmented by Literature

(Adapted from Fox (2009, p. 39-40)

<table>
<thead>
<tr>
<th>Reading (Abbott, 2007; Banerjee &amp; Wall, 2006; Jang, 2005)</th>
<th>Listening (Buck &amp; Tatsuoka, 1998, p. 141-143)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1. Skim text for overview</strong> (Abbott; Banerjee &amp; Wall; Jang)</td>
<td><strong>1. The ability to make text-based inferences, when all, or part, of the necessary information is repeated, and the response requires more than just one word</strong></td>
</tr>
<tr>
<td>CAEL specification: Getting the gist</td>
<td>CAEL specification: Getting the gist</td>
</tr>
<tr>
<td>• Identify literal meaning of main points (Jang)</td>
<td>• A low-level text-based inference was necessary to answer the item</td>
</tr>
<tr>
<td>• Use clues from the question to identify main points (Abbott)</td>
<td>• All, or part, of the necessary information was repeated</td>
</tr>
<tr>
<td>• Discriminate major arguments from supporting details (Jang)</td>
<td>• The response required more than just one word</td>
</tr>
<tr>
<td>• Skim test, identify major points, summarize main context (Abbott)</td>
<td></td>
</tr>
<tr>
<td><strong>2. Word Processing</strong> (Abbott; Jang)</td>
<td><strong>2. The ability to identify the task by determining what type of information to search for in order to complete the task</strong></td>
</tr>
<tr>
<td>CAEL specification: Comprehending vocabulary in context</td>
<td>CAEL specification: Catching specific information</td>
</tr>
<tr>
<td>CAEL specification: Identify synonyms and/or definitions</td>
<td>• The type of information required was not very obvious from just reading the question</td>
</tr>
<tr>
<td>• Break words down into smaller parts (Abbott)</td>
<td></td>
</tr>
<tr>
<td>• Identify and formulate synonyms for specific words/phrases (Abbott)</td>
<td></td>
</tr>
<tr>
<td>• Use local context to deduce word meaning (Jang)</td>
<td><strong>3. The ability to incorporate background knowledge into text processing</strong></td>
</tr>
<tr>
<td>• Use local context to deduce word meaning (Abbott)</td>
<td>CAEL specification: Diagrams, labels, charts</td>
</tr>
<tr>
<td></td>
<td>• Background knowledge would help the test taker to infer the answer</td>
</tr>
<tr>
<td><strong>3. Follow arguments throughout text, broader reading</strong> (Jang; Banerjee &amp; Wall; Abbott)</td>
<td><strong>4. The ability to process dense information, when the response requires more than just one word</strong></td>
</tr>
<tr>
<td>CAEL specification: Locating supporting info</td>
<td>CAEL specification: Paraphrasing</td>
</tr>
<tr>
<td>CAEL specification: Follow sequences of events</td>
<td>• The response required more that just one word</td>
</tr>
<tr>
<td>CAEL specification: Identify logical relations</td>
<td>• The idea units surrounding the necessary information had at least a medium amount of information</td>
</tr>
<tr>
<td>• Apply knowledge of grammar and lexical linking devices (Jang)</td>
<td></td>
</tr>
<tr>
<td>• Synthesize material from multiple sources or parts of text (Banerjee &amp; Wall)</td>
<td></td>
</tr>
<tr>
<td>• Form connections between parts of text, synthesize scattered information (Abbott)</td>
<td><strong>5. The ability to use previous items to help information location, when all, or part, of the necessary information is repeated, and the response requires more than just one word.</strong></td>
</tr>
<tr>
<td>• Read widely but selectively to recognize implicit information (Jang)</td>
<td>CAEL specification: Understanding the main idea</td>
</tr>
<tr>
<td></td>
<td>• There was a previous item on the same text</td>
</tr>
<tr>
<td></td>
<td>• All, or part, of the necessary information was repeated</td>
</tr>
<tr>
<td></td>
<td>• Crucial words for the answer had heavy stress</td>
</tr>
<tr>
<td>Reading (Abbott, 2007; Banerjee &amp; Wall, 2006; Jang, 2005)</td>
<td>Listening (Buck &amp; Tatsuoka, 1998, p. 141-143)</td>
</tr>
<tr>
<td>--------------------------------------------------------</td>
<td>------------------------------------------------</td>
</tr>
<tr>
<td><strong>4. Addressing the question</strong> (Abbott; Banerjee &amp; Wall)</td>
<td><strong>6. The ability to identify relevant information without any explicit marker to indicate that, when all, or part, of the necessary information is repeated.</strong></td>
</tr>
<tr>
<td>CAEL specification: Identifying Main idea</td>
<td>CAEL specification: Distinguishing relevant and irrelevant information</td>
</tr>
<tr>
<td>• Match key vocabulary in text and question (Abbott)</td>
<td>• The ability to recognize and use redundant information</td>
</tr>
<tr>
<td>• Apply and understand information in text (Banerjee &amp; Wall)</td>
<td>• The ability to identify relevant information with any explicit marker to indicate that</td>
</tr>
<tr>
<td>• Use the discourse format/organization to answer the question (Abbott)</td>
<td></td>
</tr>
<tr>
<td><strong>5. Ability to cope with heavy reading load</strong> (Banerjee &amp; Wall)</td>
<td><strong>7. The ability to process L2 concepts which have no literal equivalence in L1</strong></td>
</tr>
<tr>
<td>CAEL specification: speeded pressure to read two-three academic texts related to a topic</td>
<td>CAEL specification: Guessing vocabulary from context</td>
</tr>
<tr>
<td>• A concept in the necessary information was expressed differently in English from how it would be expressed in the L2</td>
<td></td>
</tr>
<tr>
<td><strong>8. The ability to incorporate background knowledge into text processing</strong></td>
<td><strong>8. The ability to incorporate background knowledge into text processing</strong></td>
</tr>
<tr>
<td>CAEL specification: Making connections to other material</td>
<td>CAEL specification: Making connections to other material</td>
</tr>
<tr>
<td>• Background knowledge would help the test taker infer the answer</td>
<td></td>
</tr>
</tbody>
</table>

Note: The italicized phrase at the top of each cell indicates the subcomponent and corresponding test specification(s).
Appendix C: Reading, Listening, and Writing Skill Subcomponents

Reading Skill Subcomponents in Relation to the Pilot Specifications

<table>
<thead>
<tr>
<th>Skill subcomponent</th>
<th>Reading One Questions</th>
<th>Reading Two Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Getting the gist</td>
<td>13, 15</td>
<td>4, 16</td>
</tr>
<tr>
<td>2. Main Idea</td>
<td>2, 5</td>
<td>2, 7, 9, 14, 16</td>
</tr>
<tr>
<td>3. Locating supporting</td>
<td>7, 9, 10</td>
<td>3, 6, 10, 13</td>
</tr>
<tr>
<td>detail</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Vocabulary Knowledge</td>
<td>3-6, 10, 11</td>
<td>5, 11, 13</td>
</tr>
</tbody>
</table>

Listening skill subcomponents in relation to the pilot specifications

<table>
<thead>
<tr>
<th>Skill subcomponent</th>
<th>Listening Questions</th>
<th>Skill subcomponent</th>
<th>Listening Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Getting the gist</td>
<td>1, 6, 7, 9, 12,</td>
<td>5. Understanding the main idea</td>
<td>4, 6, 9, 10,</td>
</tr>
<tr>
<td></td>
<td>13, 16</td>
<td></td>
<td>11, 16</td>
</tr>
<tr>
<td>• Catching specific information</td>
<td>3, 5, 8, 14, 15,</td>
<td>6. Distinguishing relevant and</td>
<td>12, 13, 16,</td>
</tr>
<tr>
<td></td>
<td>18, 20, 22</td>
<td>irrelevant information</td>
<td>19, 20</td>
</tr>
<tr>
<td>• Diagrams, labels, charts</td>
<td>4, 5, 13, 16, 20</td>
<td>7. Guessing vocabulary from context</td>
<td>6, 18, 20</td>
</tr>
<tr>
<td>• Paraphrasing</td>
<td>6, 7, 13, 16, 20</td>
<td>8. Making connections to other</td>
<td>2, 3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>material</td>
<td></td>
</tr>
</tbody>
</table>
Writing subcomponents in relation to pilot specifications (CAEL, 2008)

<table>
<thead>
<tr>
<th>General</th>
<th>Language/Grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does not understand the prompt</td>
<td>Vocabulary is informal (colloquial)</td>
</tr>
<tr>
<td>Does not meet length requirement</td>
<td>Word formation is weak (morphology)</td>
</tr>
<tr>
<td>Does not paraphrase/use own words</td>
<td>Structural fragments</td>
</tr>
<tr>
<td>Uses many quotations</td>
<td>Transitions are incorrect or missing</td>
</tr>
<tr>
<td>Has difficulty with punctuation</td>
<td>Modality use is insufficient or missing</td>
</tr>
<tr>
<td>Has difficulty with spelling</td>
<td>Noun use is inconsistent or missing</td>
</tr>
<tr>
<td>Cannot/does not reference appropriately</td>
<td>Articles are missing</td>
</tr>
<tr>
<td>Copies frequently</td>
<td>Errors with number (singular/plural)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Content</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Does not show an understanding of topic</td>
<td>Keywords from topic not used</td>
</tr>
<tr>
<td>Does not use information/support from readings</td>
<td></td>
</tr>
<tr>
<td>Does not include information/support from listening</td>
<td></td>
</tr>
<tr>
<td>Does not define or explain key terms</td>
<td></td>
</tr>
<tr>
<td>Does not have a clear opinion, claim or thesis statement</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Organization</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1st paragraph does not give the reader an overview of the writing to follow</td>
<td></td>
</tr>
<tr>
<td>1st paragraph does not state a thesis or claim</td>
<td></td>
</tr>
<tr>
<td>Body of writing is not divided into logical paragraphs or sections</td>
<td></td>
</tr>
<tr>
<td>Does not use logical transitions</td>
<td></td>
</tr>
<tr>
<td>Does not include a relevant conclusion</td>
<td></td>
</tr>
</tbody>
</table>
Appendix D: Learner Profile (Teacher Version)

Diagnostic Profile: Learner Profile

<table>
<thead>
<tr>
<th>Name of Student:</th>
<th>[Class Name]</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Skill</th>
<th>Student Score</th>
<th>Section A Mean</th>
<th>Section B Mean</th>
<th>Recommended targeted development: Include purposeful comments based on targeted needs (e.g.: “Identified as targeted reading, but could benefit from...”)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read</td>
<td>71</td>
<td>50</td>
<td>40</td>
<td>Reading [ ] Listening [ ] Writing [ ] Balanced [ ]</td>
</tr>
<tr>
<td>Listen</td>
<td>76</td>
<td>40</td>
<td>40</td>
<td>[Student name] is overall an extremely strong student based on her high listening and reading band scores. [Student name] understand the structure of writing, but is limited in her application. This could be because she does not have a deep understanding of the topic. She could benefit from strategies on how to draw on her reading notes/questions to inform her writing.</td>
</tr>
<tr>
<td>Write</td>
<td>50</td>
<td>40</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

(+ or -) accuracy of self assessment: It is clear that she underestimates her listening and reading she is accurate in her assessment of her writing. She is aware she can write and recognizes her weakness in developing an argument.

(Circle the descriptors ONLY for the skill area being targeted – circle all skills if outcome is balanced)

**Reading developmental analysis: (based on pilot specifications):**

<table>
<thead>
<tr>
<th>Objective</th>
<th>Reading One Questions</th>
<th>Reading Two Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Getting the gist</td>
<td>13, 15</td>
<td>4, 16</td>
</tr>
<tr>
<td>Main Idea</td>
<td>2, 5</td>
<td>2, 7, 9, 14, 16</td>
</tr>
<tr>
<td>Locating supporting detail</td>
<td>7, 9, 10</td>
<td>3, 6, 10, 13</td>
</tr>
<tr>
<td>Vocabulary Knowledge</td>
<td>3-6, 10, 11</td>
<td>5, 11, 13</td>
</tr>
</tbody>
</table>

**Listening developmental analysis: (based on pilot specifications):**

<table>
<thead>
<tr>
<th>Objective</th>
<th>Listening Questions</th>
<th>Objective</th>
<th>Listening Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Getting the gist</td>
<td>1, 6, 7, 9, 12, 13, 16</td>
<td>Understanding the main idea</td>
<td>4, 6, 9, 10, 11, 16</td>
</tr>
<tr>
<td>Catching specific information</td>
<td>3, 5, 8, 14, 15, 18, 20, 22</td>
<td>Distinguishing relevant and irrelevant information</td>
<td>12, 13, 16, 19, 20</td>
</tr>
<tr>
<td>Diagrams, labels, charts</td>
<td>4, 5, 13, 16, 20</td>
<td>Guessing vocabulary from context</td>
<td>6, 18, 20</td>
</tr>
<tr>
<td>Paraphrasing</td>
<td>6, 7, 13, 16, 20</td>
<td>Making connections to other material</td>
<td>2, 3</td>
</tr>
</tbody>
</table>
### Diagnostic Profile: Learner Profile

**Writing developmental trajectory (based on pilot specifications):**  
* High-level Writing Student

<table>
<thead>
<tr>
<th>General</th>
<th>Organization</th>
<th>Language / grammar</th>
</tr>
</thead>
</table>
| _Does not understand the prompt_  
_Does not meet length requirement_  
_Does not paraphrase/ use own words_  
_Uses many quotations_  
_Copies frequently_  
_Has difficulty with punctuation_  
_Has difficulty with spelling_  
√ _Can not or does not reference appropriately_ | _1^{st} paragraph does not give the reader an overview of the writing to follow_  
_1^{st} paragraph does not state a thesis or claim_  
_body of writing is not divided into logical paragraphs or sections_  
_Writer does not use logical transitions_  
_Writer does not include a relevant conclusion_ | _vocabulary is informal (colloquial)_  
_keywords from topic not used_  
_word formation is weak (morphology – word endings)_  
_word category choices and use is weak (noun/verb confusion)_  
_v Has many structural fragments_  
_Structural complexity is simple rather than complex_  
_v Transitions are incorrect or missing_  
_v text is choppy rather than cohesive_  
_Modality use is insufficient or missing_  
_Verbs use is inconsistent or missing_  
_Noun use is inconsistent or missing_  
_Subject verb agreement is inaccurate_  
_Articles are missing_  
_Language frequently causes meaning to breaks down_  
_Errors with number (singular/plural)_ |

<table>
<thead>
<tr>
<th>Content</th>
<th>Notes: The information presented in the essay is on a very general level and does not include specific information. Suggestion: Focus on strategies on how to use your reading notes to write an essay.</th>
</tr>
</thead>
</table>
| _v Does not show an Understanding of topic (not deep understanding of cause and effect)_  
_does not have a clear opinion, claim or thesis statement_  
_v Does not use Information/ support from listening_  
_does not use relevant Information/ support from reading_  
_does not include information/support from readings_  
_does not use relevant Information/support from readings_  
_does not define or Explain key terms_ | _She seems to be writing like she is speaking._  
_Suggestion: Look at how other people write._ |
Appendix E: Learner Profile (Student Version)

Learner Profile: Targeted needs

Name of Student: 

Based on the observed outcomes of your diagnostic assessment and self assessment questionnaire, your instructor has identified the following learning priorities for you this term. Please read this report carefully, and put it in your PORTFOLIO immediately.

<table>
<thead>
<tr>
<th>Skill</th>
<th>Student score</th>
<th>Class Mean</th>
<th>(+ or -) accuracy of self assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read</td>
<td>71</td>
<td>40</td>
<td>You are underestimating your reading and listening skills. For your writing, you are aware that you struggle in making a clear argument.</td>
</tr>
<tr>
<td>Listen</td>
<td>76</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Write</td>
<td>50</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

Results (take the information directly from the Learner Profile Teacher Report):

Recommended targeted area for development, explanation and associated objectives:

☐ Reading: Your diagnostic assessment learner profile indicates that you are stronger in listening and writing. As such you should work on strengthening your reading skills, especially as they pertain to getting the gist from a reading, indentifying the main idea, locating supporting details and/ or understanding vocabulary. The specific tasks you will work on for your portfolio throughout the term will be designed to target these areas. For instance, you will practice reading speed, topic vocabulary, paraphrasing and reading comprehension.

☐ Listening: Your diagnostic assessment learner profile indicates that you are stronger in reading and writing. As such you should work on strengthening your listening skills, especially as they pertain to listening for the gist of the idea, understanding the main idea, catching specific information, interpreting visuals such as graphs and diagrams, distinguishing relevant from irrelevant information, guessing vocabulary in context, and paraphrasing. The specific tasks you will work on for your portfolio throughout the term will be designed to target these areas. For instance, you will practice listening, paraphrasing, identifying signals and cues, and note-taking.
Learner Profile: Targeted needs

☑️ Writing: Your diagnostic assessment learner profile indicates that you are stronger in reading and writing. As such you should work on strengthening your writing skills, especially in the following areas:

<table>
<thead>
<tr>
<th>General</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗ Does not understand the prompt</td>
<td>✗ 1st paragraph does not give the reader an overview of the writing to follow</td>
</tr>
<tr>
<td>✗ Does not meet length requirement</td>
<td>✗ 1st paragraph does not state a thesis or claim</td>
</tr>
<tr>
<td>✗ Does not paraphrase/ use own words</td>
<td>✗ body of writing is not divided into logical paragraphs or sections</td>
</tr>
<tr>
<td>Uses many quotations</td>
<td>Writer does not use logical transitions</td>
</tr>
<tr>
<td>Copies frequently</td>
<td>Writer does not include a relevant conclusion</td>
</tr>
<tr>
<td>Has difficulty with punctuation</td>
<td></td>
</tr>
<tr>
<td>Has difficulty with spelling</td>
<td></td>
</tr>
<tr>
<td>✗ Can not or does not reference appropriately</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Content</th>
<th>Language / grammar</th>
</tr>
</thead>
<tbody>
<tr>
<td>✗ Does not show an Understanding of topic (not deep understanding of cause and effect)</td>
<td>✗ vocabulary is informal (colloquial)</td>
</tr>
<tr>
<td>✗ does not have a clear opinion, claim or thesis statement</td>
<td>✗ keywords from topic not used</td>
</tr>
<tr>
<td>✗ does not use Information/ support from listening</td>
<td>✗ word formation is weak (morphology – word endings)</td>
</tr>
<tr>
<td>✗ does not use relevant Information/ support from readings</td>
<td>✗ word category choices and use is weak (noun/verb confusion)</td>
</tr>
<tr>
<td>✗ does not include Information/ support from readings</td>
<td>✗ Has many structural fragments</td>
</tr>
<tr>
<td>✗ does not use relevant Information/ support from readings</td>
<td>✗ Structural complexity is simple rather than complex</td>
</tr>
<tr>
<td>✗ does not define or Explain key terms</td>
<td>✗ Transitions are incorrect or missing</td>
</tr>
<tr>
<td></td>
<td>✗ text is choppy rather than cohesive</td>
</tr>
<tr>
<td></td>
<td>✗ Modality use is insufficient or missing</td>
</tr>
<tr>
<td></td>
<td>✗ Verbs use is inconsistent or missing</td>
</tr>
<tr>
<td></td>
<td>✗ Noun use is inconsistent or missing</td>
</tr>
<tr>
<td></td>
<td>✗ Subject verb agreement is inaccurate</td>
</tr>
<tr>
<td></td>
<td>✗ Articles are missing</td>
</tr>
<tr>
<td></td>
<td>✗ Language frequently causes meaning to breaks down</td>
</tr>
<tr>
<td></td>
<td>✗ Errors with number (singular/plural)</td>
</tr>
</tbody>
</table>

Notes: The information presented in the essay is on a very general level and does not include specific information. Focus on strategies on how to use your reading notes to write an essay.

☑️ Balanced EAP Your diagnostic assessment learner profile indicates that you perform equally across all assessed skills: reading, listening and writing. As such you should work on strengthening these skills equally. You will work on a combination of those objective indicated above under reading, listening and writing.

It is my belief and intention that by developing an awareness and understanding of your own language strengths and weaknesses, together we can target those areas of need. I will provide you with tasks for your portfolio to work on based on your targeted group and you can use your developing skills, awareness and language strengths to work on all assigned class work and homework.
Appendix F: *Assessment Use Argument for the CAELD*

**Beneficial Consequences Assessment Claim (Assessment Claim A)**
The consequences of using the CAELD for diagnostic purposes in an EAP classroom are beneficial to the teacher and students.

- **Warrant A.1:** The CAELD is efficient to administer and score within the course schedule (Beneficence of Administration).
- **Warrant A.2:** The students have adverse experiences with taking the CAELD (Beneficence of Administration).
- **Warrant A.3:** The CAELD feedback informs the teacher’s instruction of individual students and the whole class, and informs her teaching (Beneficence for Instruction).
- **Warrant A.4:** The students use the diagnostic feedback to modify their studying of academic English (Beneficence for Student Learning).
- **Warrant A.5:** The teacher and students experience no adverse effects from the decisions made from the diagnostic feedback (Beneficence of Decisions).

**Decisions Assessment Claim (Assessment Claim B)**
The decisions made from the interpretations of the CAELD (claim C) reflect the values of the teacher and students and are equitable for all of the students.

- **Value Sensitivity**
  - **Warrant B.1:** The CAELD feedback correspond to the teacher and student's values about teaching and learning (Value Sensitivity for the Teacher and Students).
  - **Warrant B.2:** The relative seriousness of classification errors are considered in relation to the existing education values of the teacher and students (Consequences of Classification Errors).

- **Equitability**
  - **Warrant B.3:** The primary decisions made by the teacher to group the students are made by decision rules, and no other considerations are used (Equitable Decision Rules).
  - **Warrant B.4:** The students are informed about the procedures for how the decisions are made, and the decision-making process occurs in this way (Stakeholders Informed about Decision-making Process).
Interpretations Assessment Claim (Assessment Claim C)

The interpretations about students’ strengths and weaknesses in reading, listening, and reading are meaningful to the course syllabus and the teaching and learning activities in the EAP course, impartial to all groups of test takers, generalizable to subsequent learning tasks used in introductory level university courses, and relevant to and sufficient for the formative decisions made by the teacher and students.

Meaningfulness
Warrant C.1: The definition of the construct, based on academic reading, listening, and writing an argumentative essay, clearly distinguishes this construct from other constructs of language assessment. The construct of the diagnostic subcomponents clearly distinguishes them from other diagnostic subcomponents (Construct and TLU domain Definition).

Warrant C.2: The CAELD engages the student in academic reading, listening, and writing an argumentative essay (Accuracy of Inferences).

Warrant C.3: The procedures for administering the CAELD enable test takers to perform at their highest level on the assessment (Eliciting Students’ Optimal Performance).

Warrant C.4: The CAELD teacher and students’ learner profiles focuses on elements of the EAP course instruction and student learning (Utility of the Assessment Report).

Warrant C.5: The teacher and students interpret the CAELD feedback appropriately (Meaningful for Stakeholders).

Warrant C.6: The test developer clearly communicates the definition of the construct in non-technical language before the administration of the CAELD. The construct was clearly defined on the learner profile for teachers and students (The Construct was Communicated to Stakeholders)

Impartiality
Warrant C.7: The CAELD does not include response formats or content that may either favour of disfavour some students in the EAP course (Impartial Format and Content).

Warrant C.8: The CAELD does not include content that may be offensive to some test takers (Culturally Appropriate).

Warrant C.9: The students have equal access to the CAELD content and procedures, and are equally familiar with testing conditions (Equal Access to Assessment Information).

Warrant C.10: The procedures for creating the learner profile are clearly described in terms that are understandable to all test takers (Equal Access to Scoring Process).

Warrant C.11: Interpretations of the students’ CAELD feedback are equally meaningful across students from different first language backgrounds and academic disciplines (Impartial Scoring).

Generalizability
Warrant C.12: The observations made are representative of learning tasks in both the EAP course and introductory university courses (Generalization to Subsequent Learning).

Warrant C13: The criteria and procedures for recording the responses to the CAELD tasks correspond closely to those that were used by the teacher in the EAP course (Generalization of scoring to TLU tasks).

Relevance
Warrant C.14: The students’ identified strengths and weaknesses in writing are connected to instructions for how the student can improve (Meaningful Observations by the Raters).

Warrant C.15: The diagnostic feedback maps onto the teacher’s overall course planning (Relevant to Course Instruction).

Warrant C.16: The diagnostic feedback maps onto the students’ learning goals for the EAP course and university studies (Relevant to Student Learning).

Sufficiency
Warrant C.17: The teacher and students are comfortable with the diagnostic feedback for their decision-making purposes, when accounting for possible interpretation errors (Sufficiency).
Assessment Records Claim (Assessment Claim D)
The scores from the CAELD are consistent across different forms and administrations of the test, and across students from different test taker groups and academic backgrounds.

Warrant D.1: The assessment records are consistent across different groups of students (Consistency).
Appendix G: Ethics Clearance and CHRPP certificate

January 13, 2010

Ms. Christine Doe
Faculty of Education
Queen's University

GREB Ref # GEDUC-485-09
Title: “Examining the Validity of the Canadian Academic English Language (CAEL) as a Diagnostic Assessment: An Exploration of Its Test Construct, Implementation, and Viability”

Dear Ms. Doe:

The General Research Ethics Board (GREB), by means of a delegated board review, has cleared your proposal entitled “Examining the Validity of the Canadian Academic English Language (CAEL) as a Diagnostic Assessment: An Exploration of Its Test Construct, Implementation, and Viability” for ethical compliance with the Tri-Council Guidelines (TCPS) and Queen’s ethics policies. In accordance with the Tri-Council Guidelines (article D.1.6) and Senate Terms of Reference (article G), your project has been cleared for one year. At the end of each year, the GREB will ask if your project has been completed and if not, what changes have occurred or will occur in the next year.

You are reminded of your obligation to advise the GREB, with a copy to your unit REB; of any adverse events that occur during this one year period (details available on webpage www.queensu.ca/vpr/grebl/addforms.html/Adverse). An adverse event includes, but is not limited to, a complaint, a change or unexpected event that alters the level of risk for the researcher or participants or situation that requires a substantial change in approach to a participant(s). You are also advised that any adverse events must be reported to the GREB within 48 hours.

You are also reminded that all changes that might affect human participants must be cleared by the GREB. For example you must report changes in study procedures or implementations of new aspects into the study procedures on the Ethics Change Form that can be found at http://www.queensu.ca/vpr/grebl/addforms.html/Change. These changes must be sent to Linda Frid at the Office of Research Services or FRD1@queensu.ca prior to implementation. Ms. Frid will forward your request for protocol changes to the appropriate GREB reviewers and/or the GREB Chair.

On behalf of the General Research Ethics Board, I wish you continued success in your research.

Yours sincerely,

Joan Stevenson, PhD
Professor and Chair
General Research Ethics Board

copies: Dr. Malcolm Welsh, Chair Unit REB
Dr. Lying Cheng, Faculty Supervisor
E-REB: c/o Graduate Studies and Bureau of Research, Attn: Celina Pretus
Ethics Clearance Form

This is to certify that the Carleton University Research Ethics Board has examined the application for ethical clearance. The REB found the research project to meet appropriate ethical standards as outlined in the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans and, the Carleton University Policies and Procedures for the Ethical Conduct of Research.

X New clearance

Renewal of original clearance

Original date of clearance:

Date of clearance 10 June 2010
Researcher Christine Doe
Ph.D. student
Status

Department Faculty of Education, Queen’s University
Supervisor Professor Liying Cheng

Title of project Examining the Validity of the Canadian Academic English Language (CAEL) as a Diagnostic Assessment: An Exploration of its Test Construct, Implementation, and Viability

Clearance expires 31 May 2011

All researchers are governed by the following conditions:

Annual Status Report: You are required to submit an Annual Status Report to either renew clearance or close the file. Failure to submit the Annual Status Report will result in the immediate suspension of the project. Funded projects will have accounts suspended until the report is submitted and approved.

Changes to the project: Any changes to the project must be submitted to the Carleton University Research Ethics Board for approval. All changes must be approved prior to the continuance of the research.

Adverse events: Should any participant suffer adversely from their participation in the project you are required to report the matter to the Carleton University Research Ethics Board. You must submit a written record of the event and indicate what steps you have taken to resolve the situation.

Suspension or termination of clearance: Failure to conduct the research in accordance with the principles of the Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans and the Carleton University Policies and Procedures for the Ethical Conduct of Research may result in the suspension or termination of the research project.

Leslie J. MacDonald-Hicks
Research Ethics Board Coordinator
For the Chair of the Carleton University Research Ethics Board
Prof. Antonio Gueltieri
Certificate of Completion

This document certifies that

Christine D Doe

has completed the Queen's University online Course in Human Research Participant Protection (CHRPP).

Date of Issue: October 24, 2009
Appendix H: Instructions for Reviewing Essays (Phase 1)

Instructions for Reviewing Essays

While reviewing the essays, think about these questions below as you would when marking a CAEL essay that is used as a diagnostic test.

1. Turn on the recorder.

2. State the date and the essay code.

3. Read the essay aloud.

4. While you read aloud the essay, “talk aloud” about what you are thinking as you mark the essay. Specifically, “talk aloud” about:
   a. the features of writing that help place the essay in a sublevel,
   b. the features of writing that hold the essay back from being in the next sublevel, and
   c. the extra features of writing that you take into consideration in marking because the student wrote this as a diagnostic test.

6. What band score would you assign to the essay?

7. Now, imagine that this student was in front of you (as if you were his or her tutor).
   • What advice would you give him or her to reach the next band score level?
Appendix I: Rater Interview Protocol (Phase 1)

I. Perceptions of the use of CAEL for diagnostic purposes:

1. What does diagnostic assessment mean to you? [P]

2. What does diagnostic assessment look like in practice (for teachers and students)? [P]

3. What is the goal of diagnostic assessment? [P]

4. In your opinion, what is the purpose or goal of diagnosing writing for EAP students? [P]

5. What is a strength of student writing that you typically see in a CAEL essay? [PSW]
   a. Can you describe what this strength typically looks like?
   b. What is a telling sign of this type of strength?
   c. What are some other strengths?

6. What is a weakness of student writing that you typically see in a CAEL essay? [PSW]
   a. Can you describe what this weakness typically looks like?
   b. What is a telling sign of this type of weakness?
   c. What are some other weaknesses?

7. What aspects of student writing need to improve for the writing to move from a band score of 30 to 40? What about from a band score of 40 to 50? And, what about from a band score of 50 to 60? [SW]

8. In your opinion, what is the best way for students to improve their writing? [P/PSW]

II. Insight into scoring the CAEL diagnostic writing:

9. Describe for me how you approach the marking of a CAEL diagnostic essay? [AM]

10. What specific features of writing are you taking into consideration when marking a CAEL diagnostic essay? [AM]

11. Is there anything else that you are taking into consideration when marking a CAEL diagnostic essay? [AM]

12. While doing the think aloud, were there any essays that were more difficult to talk about than others? Do you remember which ones and why? [AM]

13. While doing the think aloud, were there any essays that were easier to talk about than others? Do you remember which ones and why? [AM]

14. Were there any differences between how you normally mark a CAELD essay to completing the think alouds? [AM]

III. Background Information:

15. What are your experiences with marking CAEL? [BI]
16. How long have you worked for CAEL? How long have you worked for CAEL as a rater? [BI]

17. On average, how many CAEL essays do you mark a month? [BI]

18. Do you have any other experiences with marking language tests other than CAEL? [BI]

19. What is your experience teaching ESL and EAP students? [BI]

20. What is your educational background? [BI]
Appendix J: Instructor Interview Protocol (Phase 2)

Interview 1

1. Personal Philosophy towards Teaching
   a. How would you describe your philosophy to teaching? [BI]
   b. How would you describe your philosophy towards classroom assessment? (Formative and Summative) [BI]
   c. Could you describe your experience teaching in the EAP program? [BI]
   d. What do you think is the role of the teacher in an English for Academic Purposes (EAP) course? [BI]
   e. What is the role of a student in an English for Academic Purposes (EAP) course? [BI]
   f. What do you feel is the goal of a course like [Intermediate EAP course]? [BI]

2. Orientation towards Diagnostic Assessment
   a. What does diagnostic assessment mean to you? (What is the purpose? What are the features necessary for it to be considered diagnostic assessment?) How is this different from formative or other forms of assessment? [P]
   b. Could you describe your previous experiences using diagnostic assessment? [BI]
   c. Has your orientation to diagnostic assessment changed since your first experience using such an approach? If so, how? What influenced that change? [BI]
   d. What do you think is the role of a teacher in a course that uses diagnostic assessment? [P]
   e. What do you think is the role of a student in a course that uses diagnostic assessment? [P]
   f. What is the ideal setting of using diagnostic assessment in an English for Academic Purposes (EAP) classroom? For you and/or your students? [P]

3. Considering the Diagnostic Assessment for this Term
   a. How do you plan to use the diagnostic feedback in your teaching and course planning? [DM]
b. Has the process of completing the diagnostic profiles influenced your teaching? If so, how? [DM]

c. At this point in the course, do you feel the diagnostic assessment feedback has helped improve your teaching? Why or why not? [P]

d. What surprised you about receiving the diagnostic feedback? About a particular student? The whole class? Anything else? [P]

e. What elements of the diagnostic feedback do you think are most useful for you and your teaching? Why? [P]

f. What elements of the diagnostic feedback do you think are not useful for you and your teaching? Why not? [P]

g. If you had a chance to go back and change the process of giving the diagnostic assessment, what would you do differently? Why? [DM]

h. If you had a chance to change the instrument used, what would you change? Why? [DM]
Teacher Interview 2

1. **Decision-making based on the Diagnostic Assessment**
   a. Have you integrated the diagnostic assessment feedback into your *course planning*? If yes, [DM]
   b. What decisions have you made to integrate the diagnostic assessment feedback into your *course planning*?
   c. Describe the process of integrating the diagnostic assessment feedback into your *course planning* so far?
   d. How did you arrive at these decisions?
   e. Looking back, what would you change?

4. **Have you integrated the diagnostic assessment feedback into your daily teaching so far?** If yes, [DM]
   a. What decisions have you made to integrate the diagnostic assessment feedback into your *daily teaching*?
   b. Describe the process of integrating the diagnostic assessment feedback into your *daily teaching*?
   c. How did you arrive at these decisions?
   d. Looking back, what would you change?

5. **Have you used the diagnostic assessment feedback to support your students’ learning so far?** If yes, [DM]
   a. What decisions have you made to integrate the diagnostic assessment feedback into your *students’ learning so far*?
   b. Describe the process of integrating the diagnostic assessment feedback into your *students’ learning so far*?
   c. How did you arrive at these decisions?
d. Looking back, what would you change?

6. Have you integrated the diagnostic assessment feedback into your assessment of students? If yes, [DM]
   a. What decisions have you made to integrate the diagnostic assessment feedback into your assessment of students?
   b. Describe the process of integrating the diagnostic assessment feedback into your assessment of students?
   c. How did you arrive at these decisions?
   d. Looking back, what would you change?

7. Have you integrated the diagnostic assessment feedback into the targeted groups led by others? If yes, [DM]
   a. What decisions have you made to integrate the diagnostic assessment feedback into the targeted groups led by others?
   b. Describe the process of integrating the diagnostic assessment feedback into the targeted groups led by others?
   c. How did you arrive at these decisions?
   d. Looking back, what would you change?

Using the Diagnostic Assessment

8. Course planning (Ask if yes to Question 1 above)
   a. What are the next steps for using the diagnostic assessment in your course? [DM]
   b. What has been successful about using the diagnostic feedback in planning your course?
      Is there one incident that strikes you as having been really successful that you could describe for me?
   c. What has been challenging? Is there one incident that strikes you as having been really challenging that you could describe for me?
9. **Daily teaching (Ask if yes to Question 2 above) [DM/P]**
   
a. What has been successful about using the diagnostic feedback in your daily teaching? Is there one incident that strikes you as having been really successful that you could describe for me?

   b. What has been challenging? Is there one incident that strikes you as having been really challenging that you could describe for me?

10. **Course planning (Ask if yes to Question 3 above) [DM/P]**
   
a. What has been successful about using the diagnostic feedback in students’ learning so far? Is there one incident that strikes you as having been really successful that you could describe for me?

   b. What has been challenging? Is there one incident that strikes you as having been really challenging that you could describe for me?

11. **Assessment of students (Ask if yes to Question 4 above) [DM/P]**
   
a. What has been successful about using the diagnostic feedback into your assessment of the students? Is there one incident that strikes you as having been really successful that you could describe for me?

   b. What has been challenging? Is there one incident that strikes you as having been really challenging that you could describe for me?

12. **Targeted Groups (Ask if yes to Question 5 above) [DM/P]**
   
a. What has been successful about using the diagnostic feedback into the targeted groups? Is there one incident that strikes you as having been really successful that you could describe for me?

   b. What has been challenging? Is there one incident that strikes you as having been really challenging that you could describe for me?
13. **Usefulness of Diagnostic Assessment**

   a. Overall, in using the diagnostic feedback what has not worked as well as you had hoped?  
      [P]

   b. What do you think contributed to these difficulties? [DM]

   c. Do you think there are particular students that are benefiting from the process more than others? Do you think there are any students that are not benefiting from the process? [P]

   d. What aspects of the process have surprised you? [P]

   e. If you were to do it again, what would you do differently? [DM]

   f. If you were to do it again, what would remain the same? [DM]
Appendix K: Classroom Observation Protocol (Phase 2)

Guiding Questions:

A: General Observations of Teaching (for contextual understanding):

1. What is the overall atmosphere of the classroom?
2. How are the students involved in classroom discussion?
3. How are topics introduced to the students?
4. What are the regular classroom teaching interactions?

B: Implementation of the Diagnostic Assessment Feedback:

1. How is the diagnostic feedback discussed by the teacher with the students?
2. How is the diagnostic feedback integrated with classroom teaching (i.e., are students placed into levels of groups based on diagnostic feedback)?
3. How does the teacher interact with individual students about the diagnostic feedback?
4. How does the teacher plan diagnostic feedback in teaching to support learning?

Template for organizing classroom observation notebook:

<table>
<thead>
<tr>
<th>Episode</th>
<th>Time</th>
<th>Activity</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix L: Student Reflections (Phase 3)

Reflection 1

Name: ____________________________  Section: __________________

1. How much do you think that the diagnostic information presented in the report reflects your actual language ability? Please explain and be specific.

________________________________________________________________________

________________________________________________________________________

2. How useful do you think the information presented in the report/profile will be to help you improve your English skills? Please explain and be specific.

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

2b. Circle the letter that best represents your answer above.
   a. Not at all   b. A little   c. Somewhat   d. Quite a lot   e. Completely

3. Were you surprised by any of the information presented in the diagnostic report? Please explain and be specific.

________________________________________________________________________

________________________________________________________________________

________________________________________________________________________

4. Have you taken the CAEL before taking it in this class?  1) Yes  2) No

5. How useful do you think the information would be to help you improve your score on the CAEL in the future?

________________________________________________________________________

________________________________________________________________________

6. How useful do you think the information would be to help you improve your performance in 1500 this term?

________________________________________________________________________

________________________________________________________________________
6. What skills (reading, writing, listening, etc) do you plan to improve on while studying in 1500? Please explain how you plan to improve these skills.

7b. To what degree was your plan (the question above) influenced by the diagnostic report that you received?
   a. Not at all   b. A little   c. Somewhat   d. Quite a lot   e. Completely

7c. What else influenced your plan of improving your English skills? Please explain and be specific.

8. How do you think the process could be improved? (Taking the test, receiving the feedback)

9. How important is it that the diagnostic feedback considered the three skills of reading, writing, and listening?
Reflection 2

Name: __________________________________ Section: ____________________

1. Please indicate how often you have thought about the diagnostic feedback (learner profile) when studying on your own and outside of [the Intermediate EAP course]?

   a. Never   b. Once or twice   c. Once a week   d. More than once a week

1b. If you answered (a), explain why not? If you answered (b, c, or d), explain what you were doing and how you used the diagnostic feedback?

2. Please indicate how often you have thought about the diagnostic feedback (learner profile) when studying in [the Intermediate EAP course]?

   a. Never   c. Once or twice   d. Once a week   e. Every class

2b. If you answered (a), explain why not? If you answered (b, c, or d), explain what you were doing and how you used the diagnostic feedback?

3. List the skill(s) that you had indicated on the last reflection that you planned to improve.

3b. What have you done to improve the skill(s)? Please explain and be specific by giving an example. (what skills and how).

4. What skill or skills do you plan to improve on now? ________________________________

4b. Please explain your next steps of how you plan to improve the skill or skills before the end of term. (consider strategies that you are learning in class and in your skills groups)
4c. To what degree was your plan (Question 4b) influenced by the diagnostic report that you received?

   a. Not at all   b. A little   c. Somewhat   d. Quite a lot   e. Completely

4d. What else influenced your plan (Question 4b) of improving your English skills (e.g., test scores, feedback from [teacher’s name], targeted teacher, my friends, my knowledge or experience)?

5. So far, how useful was the diagnostic report in improving your academic English skills?

   a. Not at all   b. A little   c. Somewhat   d. Quite a lot   e. Completely
Reflection 3

Name: ___________________________ Section: ________

1. Please indicate how often you have thought about the diagnostic feedback (learner profile) since the last reflection, when studying on your own and outside of [the Intermediate EAP course]?
   
   a. Never   b. Once or twice   c. Once a week   d. More than once a week

1b. If you answered (a), explain why not? If you answered (b, c, or d), explain what you were doing and you how used the diagnostic feedback?

2. Please indicate how often you have thought about the diagnostic feedback (learner profile) since the last reflection, when studying in [the Intermediate EAP course]?

   a. Never   b. Once or twice   c. Once a week   d. Every class

2b. If you answered (a), explain why not? If you answered (b, c, or d), explain what you were doing and you how used the diagnostic feedback?

3. List the skill(s) that you had indicated on the last reflection that you planned to improve.

3b. What have you done to improve the skill(s)? Please explain and be specific by giving an example. (what skills and how).

4. What skill or skills do you plan to improve on now? ________________________________

4b. Please explain your next steps of how you plan to improve these skills now that the course is finished.

4c. To what degree was your plan (Question 4b) influenced by the diagnostic report that you received?
   
   a. Not at all   b. A little   c. Somewhat   d. Quite a lot   e. Completely

4d. What else influenced your plan (Question 4b)? For example, test scores, feedback from [name of teacher], your own knowledge.
5. If it were possible, would you like to take the diagnostic test (CAELD) again?
   a. Yes               b. No

5b. Why or Why Not? Please explain and be specific.

5c. If you did take the diagnostic test (CAELD) again, would you expect the results of your diagnostic report to be different compared to when you took it in January?
   a. Yes               b. No

5d. What would the feedback on your diagnostic report say your strengths and weaknesses are now? Which skills would be targeted for improvement?

6. How important is it that a course, like [Intermediate EAP], use diagnostic assessment?
   a. Not at all   b. A little   c. Somewhat   d. Quite a lot   e. Completely

6b. Please explain and be specific.

7. Overall, how useful was the diagnostic report in improving your academic English skills?
   a. Not at all   b. A little   c. Somewhat   d. Quite a lot   e. Completely

8. Please fill in the background information for Christine’s Research:

Gender ___ (M/F) Age: ____ Home Country: ___________ First Language: ______________
Appendix M: Students Interview Protocol (Phase 3)

Interview Questions for Students (Interview 1)

1. Background Information
   a. What is your name?
   b. Where were you born?
   c. What is your home country?
   d. What is the language you first spoke at home? Do you speak any other languages?
   e. Do you mind me asking, how old are you?
   f. When did you come to Canada? Have you lived in any other English speaking countries other than Canada?
   g. How long have you been studying English for? Have you ever taken other courses like 1500 before?
   h. Did you take a language test to get into the EAP program? Which one? Do you remember your score?
   i. Had you ever heard of the term “diagnostic assessment” before this course?
   j. Had you ever experienced this type of approach to teaching before this course?

2. Reactions to the Diagnostic Assessment
   a. When you took the CAEL, you knew that it was for the diagnostic assessment and the results would not matter for your marks, right? How did you feel while taking the CAEL? Can you describe your feeling while taking the CAEL? Were you nervous?
   b. What was your score on the CAEL diagnostic? What feedback did you get?
c. What was the first thing that you did when you received the learner profile? What next?

d. Were surprised with any of the diagnostic feedback?

e. What did you like about the profile?

f. What did you not like about the profile?

3. Using the Diagnostic Information

a. Have you thought about the diagnostic feedback while studying English at home or in class? If so, what are some examples of how?

b. Did you change any of your studying habits because of the diagnostic feedback? If so, which ones?

c. What parts of the diagnostic feedback were most useful for you and your studying? Why? [With an example diagnostic profile in front of the student]

d. What parts of the diagnostic feedback were not useful for you and your studying? Why not? [With an example diagnostic profile in front of the student]

e. Do you think that having the diagnostic assessment included in the course makes improves the course? If so, how?

   (For your learning? To Diane’s teaching? Does it make it more interesting?)

f. If you had to go through the process again, what would you want to change?

g. Do you feel that the CAEL was a good test for showing your language ability? What about for finding out about your strengths and weaknesses?

h. If you take an EAP class again, would you like it if the course included diagnostic assessment?
Interview 2 – About Use of Diagnostic Assessment

I. Use of diagnostic feedback

1. First, I would like you to put a check mark (✓) next to the activities that you think helped the diagnostic assessment process (taking the test, getting the feedback, using the feedback). Write in anything else that you can think of in the blanks provided. [C]

- Self
- Diane’s Teaching
- Portfolio Teaching
- Reflections

  Learner’s Profile (diagnostic report)

- __________________________
- __________________________
- __________________________
- __________________________

2. If the interviewee does not check *Self*, ask why? Then, ask how they think they would have answered the following questions.

Or,

3. Have you thought about or used the diagnostic feedback since our last interview? If so, how?

4. How did you connect the diagnostic feedback to your studying?

5. What was useful for helping you connect the feedback to your studying? What would have helped you use the feedback better?
Teacher’s Use

6. If the interviewee does not check Diane’s teaching, ask why?

Or,

7. How do you feel Diane’s teaching has helped the diagnostic process?

8. Do you think that Diane connected the diagnostic feedback to her teaching? If so, how?
   [possible follow-up question]

9. What aspects of Diane connecting the diagnostic feedback to her teaching was useful?
   What do you think would have made the process more successful for your learning?

Portfolio Teaching

10. If the interviewee does not check Portfolio Teaching, ask why?

Or,

11. How do you feel the Portfolio Teaching has helped the diagnostic process?

12. What small group are you in? How did the portfolio teacher connect the diagnostic feedback to the activities given?

13. What was useful by having the portfolio teacher connect the feedback to the activities?
   What do you think would have made the process more successful for your learning?

Learner Profile

14. If the interviewee does not check Learner Profile, ask why?

Or,

15. How do you feel the Learner Profile has helped the diagnostic process?

16. Since our last interview, have you looked at the learner profile? If so, how and when?

17. When looking at this feedback what aspects were most useful for you in trying to improve? Is there anything missing?
Reflections

18. If the interviewee does not check Reflections, ask why?

Or,

19. How do you feel the Reflections has helped the diagnostic process?

Advice

20. Now that you have experience the diagnostic assessment, what advice would you give a student in a course that has diagnostic assessment?

21. Overall, what was the most useful part about having the diagnostic assessment process?

22. Overall, what was the least useful part about having the diagnostic assessment process?

II. Effect of diagnostic assessment

23. Would you say that the diagnostic assessment has had an effect on your learning of English? If yes, in what ways?

24. Do you think you have improved your English skills since the beginning of the term?

   How have you improved? Do you think the diagnostic feedback is still appropriate?

   [While looking at the learner profile]

25. Overall, do you think it was useful to have the diagnostic assessment in the course?

   Would you take such a course again?

III. Perceptions of diagnostic assessment

26. What does diagnostic assessment mean to you? [P]

27. What does diagnostic assessment look like in practice (for teachers and students)? [P]

28. What is the goal of diagnostic assessment? [P]

29. Is this goal different for teachers and students? If so, how? [P]
## Appendix N: Features of Writing Subcodes

<table>
<thead>
<tr>
<th><strong>Language</strong></th>
<th><strong>Details</strong></th>
<th><strong>Organization</strong></th>
<th><strong>Details</strong></th>
<th><strong>Content</strong></th>
<th><strong>Details</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grammar</strong></td>
<td>Appropriate word order; clause structure.</td>
<td><strong>Structure</strong></td>
<td>The text contains the key pieces of an essay: introduction, thesis statement, etc.</td>
<td><strong>Addressing the topic</strong></td>
<td>The text answers the question and remains on topic</td>
</tr>
<tr>
<td><strong>Academic language use</strong></td>
<td>Appropriate use of academic conventions of writing.</td>
<td><strong>Argument</strong></td>
<td>The essay includes a logical argument</td>
<td><strong>Knowledge</strong></td>
<td>The text demonstrates that the student understood the readings and listening</td>
</tr>
<tr>
<td><strong>Tense</strong></td>
<td>Appropriate tense, typically subject-verb agreement.</td>
<td><strong>Cohesiveness</strong></td>
<td>The text is predictable and hangs together</td>
<td><strong>Support</strong></td>
<td>The student has used appropriate information from the readings or listening to support their argument</td>
</tr>
<tr>
<td><strong>Vocabulary</strong></td>
<td>Correct word usage</td>
<td><strong>Linking Ideas</strong></td>
<td>Ideas are connected together using transitioning words or phrases</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Morphology</strong></td>
<td>Inappropriate word form, e.g., verb instead of the noun</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Clarity</strong></td>
<td>The text is easily understood or not</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>Whether the student has control of the language</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Flexibility and Range</strong></td>
<td>Overall, the text contains a variety of sentence structures, range of vocabulary, etc</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>