2006-06-29

Validating the Rating Process of an English as a Second Language Writing Portfolio Exam

Robb Mark McCollum
Brigham Young University - Provo

VALIDATING THE RATING PROCESS OF AN ENGLISH AS A SECOND LANGUAGE WRITING PORTFOLIO EXAM

by

Robb Mark McCollum

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Master of Arts

Department of Linguistics and English Language

Brigham Young University

2006

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Robb Mark McCollum

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

_____	_____
Date	Diane Strong-Krause, Chair
_____	_____
Date	Neil J. Anderson
_____	_____
Date	Wendy Baker

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Robb Mark McCollum in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

_____ _____

Date

Diane Strong-Krause
Chair, Graduate Committee

Accepted for the Department

_____

Lynn E. Henrichsen
Department Chair

Accepted for the College

_____

John Rosenberg
Dean, College of Humanities

ABSTRACT

VALIDATING THE RATING PROCESS OF AN
ENGLISH AS A SECOND LANGUAGE
WRITING PORTFOLIO EXAM

Robb Mark McCollum
Department of Linguistics and English Language
Master of Arts

A validity study can be used to investigate the effectiveness of an exam and reveal both its strengths and weaknesses. This study concerns an investigation of the writing portfolio Level Achievement Test (LAT) at the English Language Center (ELC) of Brigham Young University (BYU). The writing portfolios of 251 students at five proficiency levels were rated by 11 raters. Writing portfolios consisted of two coursework essays, a self-reflection assignment, and a 30-minute timed essay. Quantitative methods included an analysis with Many-Facet Rasch Model (MFRM) software, called FACETS, which looked for anomalies in levels, classes, examinees, raters, writing criteria, and the rating scale categories. Qualitative methods involved a rater survey, rater Think Aloud Protocols (TAPs), and rater interviews.

Results indicated that the exam has a high degree of validity based on the MFRM analysis. The survey and TAPs revealed that although raters follow a similar pattern for rating portfolios, they differed both in the time they took to rater portfolios and in the

degree to which they favored the rating criteria. This may explain some of the discrepancies in the MFRM rater analysis. Conclusions from the MFRM analysis, surveys, TAPs, and interviews were all used to make recommendations to improve the rating process of the LAT, as well as to strengthen the relationship between LAT rating and classroom teaching and grading.

ACKNOWLEDGMENTS

My gratitude is extended to many individuals for their help with this research project. Dr. Diane Strong-Krause, Dr. Wendy Baker, and Dr. Neil J. Anderson have been exemplary models as instructors, mentors, and members of my thesis committee. I would also like to thank all the faculty and staff of the BYU Linguistics and English Language department; I have thoroughly enjoyed studying in this program. Additionally, I would like to thank the faculty and staff of the BYU English Language Center; I value my education in graduate courses as much as my education as a member of the ELC community. I would especially like to thank the teachers in the writing program: the writing coordinator, Nancy Tarawhiti, supported every aspect of this research, and this study would not have been possible without the cooperation of the teacher-raters. Finally, I am grateful for support of my parents, Sam and Jeannine, and for their examples of dedication, faith, and hard work; those are commodities that outweigh any monetary contribution.

Table of Contents

List of Tables

Table of Figures

CHAPTER ONE

Introduction

Rationale for This Study

The English Language Center (ELC) at Brigham Young University (BYU) provides English as a Second Language (ESL) instruction to students from around the world. As part of the ELC's attempts to improve language instruction and assessment, the Center now uses Level Achievement Tests (LATs) as the final exams at the end of each semester. The LATs are standardized for all classes at a given level (1 through 5) and skill area (grammar, listening, speaking, reading, and writing). This standardization was designed to create greater assessment uniformity across all ELC classes and to ensure that students are promoted based on language proficiency and not based on good citizenship or time spent in the program. In the past year, efforts have been made to assess whether the LATs are accomplishing their purpose. In particular, faculty and graduate students have investigated both the validity and reliability of the listening and the speaking exams (Lee, 2005; Tai, 2004).

These investigations have attempted to answer the broad question: To what degree is the test meeting the assessment and instructional goals for which it was designed? For example, the researchers have studied how the exams match the course objectives, how the exams reflect classroom teaching, how reliable the rubric and raters are, and how well the tests distinguish among examinees. These studies have had a washback effect, meaning that administrators and instructors at the ELC have been able to use the research results to improve both exam procedures and classroom teaching.

With the success of validity studies in the listening and the speaking LATs, the ELC faculty and staff are now interested in conducting similar studies for the other skill area LATs. This paper outlines an initial investigation into a validity study of the writing LAT. The same general question that has been asked of the listening and the speaking LATs will now be asked regarding the writing LAT: To what degree is the writing test meeting the assessment and instructional goals for which it was designed? In other words, how valid is the ELC's writing LAT?

Messick (1992) provides an oft-quoted definition for validity. He explains that "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (p. 1487). The purpose of a validation study is to gather various types of evidence to help assess the degree to which an instrument is functioning as expected, and also the degree to which the results are being properly interpreted and used. In order to make this evaluation, a researcher collects evidence from various aspects of the testing instrument and its applied context. It is not sufficient to make a claim of validity based on one single type of evidence; rather, a researcher will seek out numerous forms of support to establish to what degree – and not a definitive statement that – test results are valid. By necessity, a well-grounded validation study incorporates a myriad of closely-related research questions in order to make a statement about an instrument's degree of validity.

Therefore, due to the nature of validity, it is not feasible within the timeframe of this current study to attempt a broad and extensive validation study of the writing LAT. Instead, the following inquiry focuses on one aspect of the writing LAT – the rating

process – and attempts to gather several sources of validity-related evidence in regards to this particular feature of the writing LAT. Due to the highly specialized nature of this study, and the low population samples, this results of this research are not intended to be generalized beyond the immediate context. However, this study can serve as a model for portfolio assessment evaluation. Other second language education programs can benefit from this model process of gathering and analyzing both qualitative and quantitative data, and then reforming a writing exam as described in this study.

Purpose of This Study

This study was conducted to gain scoring-related validity evidence regarding the rating process used to evaluate the end-of-semester writing LAT at the ELC of BYU. This study follows an argument-based approach to validity: the validity of the writing LAT will be based on discrediting or confirming the unwanted presence of misfitting data in the Many-Facet Rasch Model analysis.

In addition, this study will attempt to contextualize and interpret the quantitative analysis with the help of rater surveys, Think Aloud Protocols, and interviews. The qualitative evidence is needed in order to better understand patterns or discrepancies in the quantitative results. Based on the responses to the quantitative and qualitative data, conclusions are drawn regarding the validity of the LAT. The ELC is interested in improving the quality of LATs, and this study will gather data regarding the evaluation of the writing exams that will later be used to improve both the teaching and the evaluating of future writing students.

Research Questions

The following questions guide the collection of data regarding the scoring-related validity of the rating process of the writing LATs. These questions are based directly on the quantitative arguments and qualitative inquiries described above.

Quantitative evidence:

1. How valid are the writing LAT scores based on a Many-Facet Rasch Model analysis? This question is further separated into the following questions.
    - How well do the LATs distinguish among levels, classes, and examinees?
    - How severe are the raters in relation to one another?
    - How consistent are the raters?
    - How difficult are the writing criteria in relation to one another?
    - How well is the rating scale used?

Qualitative evidence:

2. What is the degree of rater agreement on the priority of rating criteria among and between levels?
3. How do raters apply the rating criteria to determine a LAT score? (i.e. prioritizing some criteria more than others, and valuing some criteria more in higher levels than in lower levels, etc.)
4. How do raters use portfolio samples to negotiate a LAT score? (i.e. holistically based on all scores, or an average based on individually determined scores for each sample, etc.)

Definition of Key Terms

The following definitions will aid the reader in understanding how the following key terms will be used in this paper. They are listed below in alphabetical order:

1. *Achievement test* refers to a criterion- (and not norm-) referenced end-of-semester exam designed to evaluate students' mastery of a particular set of course objectives.
2. *Logit* is a measurement based on logarithms and item response theory. Many-facet Rasch measurement analysis uses logits to measure the relative ability (or severity, difficulty, etc.) of analyzed data.
3. *Misfit* refers to items from the many-facet Rasch measurement analysis that vary beyond the degree to which the software can compensate for their inconsistent behavior.
4. *Misorder* refers to items from the many-facet Rasch measurement analysis that appear higher (or lower) on the trait scale that expected due to their pre-assigned level or class.
5. *Rating criteria* are the detailed set of writing construct-related competencies that raters use to evaluate students' written work.
6. *Rating scale* refers to the grading categories into which students are assigned by raters based on the degree to which students meet the rating criteria.
7. *Validity* concerns the degree to which a test meets the purpose of its design and the extent to which its interpretation and application are aptly used.
8. *Writing* refers to the language skill of written composition. It is far more than grammatical ability, and also includes both micro- and macro-linguistic

features as well as the cognitive and pragmatic competence required to create a meaningful text that is aware of audience, purpose, and composition process.

Delimitations of This Study

As stated earlier, an extensive and in-depth validation study is beyond the scope of this project. Instead, this paper focuses on an evaluation of the writing LAT rating process. As such, the following aspects are delimitations of this study:

This study does not address the administration of the exam. The conditions under which the LAT is delivered to students will not be considered.

There is only a minimal investigation into the feedback and washback effects of the LAT. In other words, there is minimal investigation into the influence of LAT results on student ability, classroom teaching, and test administration after the LAT is delivered each semester.

Although it would be helpful in a validation study of this nature, this investigation does not fully address the degree to which the LAT tasks succeed in differentiating between students of varying level proficiencies. This is because there is currently no overlap in LAT tasks between skill levels; thus only minimal comparison can be made between students in various skill levels beyond the analysis performed through basic Many-Facet Rasch Model measurement.

Finally, for the purposes of this study, it is assumed that the populations (students, teachers, and raters) from one semester to another are similar, if not equal. Obviously this is not true; however, given the large groups and the overall consistency in general LAT performance across semesters, any major differences between time populations will be

assumed to be negligible. Even still, any generalizations made to extend the results of this study to future semesters at the ELC should be made with caution.

CHAPTER TWO

Literature Review

Chapter One outlined the need for a validity study of the writing Level Achievement Test (LAT) at the English Language Center (ELC) of Brigham Young University (BYU). An introduction was given to the concepts of validity and the role of gathering validity-related evidence to help verify the effectiveness of a given test. As stated in Chapter One, the quest for test validation is potentially endless; no researcher can ever claim that a given test is completely valid or invalid. Rather, a researcher can evaluate tests according to degrees of validity based on numerous sources of validity-related evidence. This validity study focuses primarily on the writing LAT's rating process.

The purpose of Chapter Two is to provide a theoretical basis for conducting such an investigation. This chapter is divided into three primary sections:

Section One – Writing Assessment: a definition of the writing skill construct is given and then several purposes to writing assessment are explained. This is followed by a discussion regarding several methods for assessing writing which leads into an explanation of writing portfolio assessment, the chosen method of assessment at the ELC. Specific benefits and problems with portfolio assessment are evaluated, including its strengths and weaknesses as a valid form of assessment.

Section Two – A Validation Study: the details of and needs for a validity study are described. Types of validation evidence are briefly described as well as methods for gathering validity-related data.

Section Three – Scoring-related Validity: the discussion of validity will be continued with an emphasis on scoring-related validity which will serve as the primary source of validation evidence for this study. Reliability, a concept pivotal to scoring-related validity, will also be discussed as it relates to both raters of performance tests and test rating scales and criteria. Several reliability studies will be cited for their ability to offer insight into both quantitative and qualitative methods of investigating validity.

Writing Assessment

For the purpose of this study, the definition of writing involves more than grammatical competence; it also involves the use of micro-linguistic features (such as spelling, grammar, punctuation, etc.) and macro-linguistic features (organization, discourse continuity, content, etc.). Although ESL composition courses may necessitate a greater emphasis on micro-linguistics features than courses for native speakers, effective writing courses will help students develop both bottom-up (micro-linguistic) and top-down (macro-linguistic) writing skills. In addition to linguistic features, effective writing also employs literary and rhetorical devices, and has a sense of audience, voice, purpose, and process. It is this multi-layered approach to composition that will serve as the construct for writing in this study.

Purposes and Types of Language Tests

The purpose of any academic test is to measure some construct in order to say something meaningful about the ability or knowledge of the test taker, which is then often used as a predictor of propensity to perform in a future situation. In the language teaching field, numerous tests are used for a variety of purposes. This includes placement tests, diagnostic tests, and achievement tests.

Placement tests are administered before students are granted entry into a particular language class. In order to decide which level the student is best suited for, administrators will require a student to take a placement test. A good placement test will group students of similar ability in the same class so that a teacher can provide instruction that will be equally beneficial to all the students (Hughes, 2003). Related to placement tests are diagnostic tests. Diagnostic tests are usually given at the beginning of a term of instruction, where administrators or researchers hope to assess a test taker's base ability before proceeding with instruction or treatment. Diagnostic tests tell the instructor about the strengths and weaknesses of the students so that the course instruction can focus primarily on helping students overcome their weaknesses (Brown & Hudson, 2002).

On the other end of the spectrum are achievement tests which are usually delivered at the end of a term of instruction. Achievement tests are designed to measure the degree to which the test taker has accomplished a course's goals or objectives. From an achievement test score, inferences are made regarding the skill-level of the test takers and the degree to which they are prepared to graduate, whether it be into a subsequent course or level, or into a profession. Hughes (2003) emphasizes the importance of relating achievement tests to course objectives in language programs. He describes the connection that language educators should strive for among course objectives, classroom teaching, exam tasks/items, and the desired skills and abilities that students will need upon completion of the course in preparation for employment or further education. An effective achievement test, such as the ELC's writing LAT, will not only meet course objectives and instruction, but will provide meaningful insight into test takers' mastery of the current level's competencies and their readiness for the next language skill level.

Indirect versus Direct Testing

Once test administrators have established at what stage within the learning process they wish to administer the test, they must then decide how to best assess the target language skills. It has long been the consensus in the educational community that writing is a language skill that is best tested directly, rather than indirectly (Hughes 2003; Jacobs et al. 1981). Indirect tests of writing tend to focus on discrete points of either mirco- or macro-linguistic knowledge, such as sentence structure, vocabulary, paragraph organization, etc.

Indirect testing is preferred for its ability to help instructors test individual aspects of language related to writing. Also, this form of assessment can easily measure points of student error such that test evaluators can provide feedback to students on specific problem areas. However, indirect testing tends only to measure knowledge about composition and error correction, but does not actually measure student ability to write a coherent piece of writing. By definition, indirect tests of composition do not directly measure writing ability. Criticism has been laid against indirect testing of writing proficiency due to concerns regarding the supposed lack of relation between indirect tests of writing and students' ability to compose good writing. In other words, just because students know a great deal about writing does not mean that they can write effectively. In consequence, it has become uncommon for language educators to use indirect testing in writing assessment. In fact, in much of their published work, many prominent writing researchers fail to even discuss forms of indirect testing of writing since the method is no longer considered an accurate measure of true writing ability for anything above extreme

beginners (Brown & Hudson, 2002; Hamp-Lyons, 1994; Hughes, 2003). Instead, current literature into writing assessment focuses on forms of direct assessment.

Direct composition tests require students to produce actual writing. These tests usually involve a prompt that places the students in a particular situation where they are required to use writing language to accomplish specific written communication objectives. Because direct tests are less focused on discrete points of knowledge, assessment is more inductive. Evaluators must infer student knowledge and understanding based on their ability to produce a meaningful piece of writing; this is the exact opposite of indirect testing where evaluators attempt to infer real writing ability from student knowledge about discrete points of writing. Direct testing is more reflective of real writing ability, but its means of evaluation are more subjective (Brown & Hudson, 2002; Hughes, 2003). In short, it is a compromise test administrators make; in order to create a test that is more reflective of real language ability, they design a test that is unavoidably less objective and more complicated to grade.

Despite this, Hughes (2003) explains that direct tests are preferable not only because they give a more accurate picture of writing ability, but also because the washback effects of direct tests are far more desirable than with indirect tests. Washback, sometimes called backwash, refers to the influence that testing has on curriculum and teaching. A test with positive washback encourages students and teachers to focus on course objectives since the exam adequately reflects and assesses the skills and knowledge taught in the course. On the other hand, negative washback can occur when there is a discrepancy between the test items and the classroom teaching. Students within a program will notice this discrepancy and will be less desirous to study the course as it is

designed. As a result, teachers will feel pressure to teach to the test, and, in turn, the curriculum will indirectly be revamped. In order to avoid negative washback effects, test designers should relate exam tasks to curriculum performance objectives, thus ensuring that teaching and assessment both focus on the same desired set of skills (Brown & Hudson, 2002).

Because indirect tests of writing tend to focus on discrete features of composition (whether macro- or micro-linguistic), they segregate assessment from the overall desired language skills and produce negative washback. Instead of helping students develop writing competency, a composition course with indirect assessment will morph into instruction *about* writing rather than a writing course. In contrast, direct tests of writing ability are perceived to have positive washback.

Performance-based assessment, a form of direct testing, encourages the use of authentic language skills. Performance tests require students to complete tasks that are reflective of real-world language situations that students will encounter outside of the classroom. The timed essay, perhaps the most common performance test in writing assessment, requires test takers to create a composition based on a written prompt. The type of prompt may vary so as to encourage a variety of different writing from test takers. For example, administrators hoping to assess business written communication skills might give test takers a scenario requiring them to write a business letter. Another performance test may be interested in assessing academic ability; the prompt for such a test may ask students to write an opinion essay based on a controversial written statement provided to the examinee. Performance-based assessment allows evaluators to judge how test takers combine discrete points of writing knowledge into a cohesive composition that

employs both writing theory and practical situations. However, there are numerous problems associated with the common timed essay performance test. As such, language educators have begun to adopt a new form of performance-based assessment – the writing portfolio.

Writing Portfolio Assessment

Over the last decade, writing portfolios have gained increasing popularity as a new form of performance-based writing assessment (Hamp-Lyons & Condon, 2002; Huot, 2002; Hyland, 2002; Weir, 2005). Because portfolios are a fairly new form of writing assessment and because they do not conform to traditional test-taking procedures, portfolios are frequently categorized as an "alternative form of assessment" (Bailey, 1998; Brown & Hudson, 2002; Hughes, 2003). Although there is no consensus on exactly what constitutes or defines "alternative assessment," researchers tend to agree that alternative assessment involves testing methods that are easily incorporated into classroom instruction and allow for greater student involvement in the test material selection. Coombe and Barlow (2004) provide the following criteria for alternative assessment:

- Emphasis on assessing individual growth rather than comparisons with peers.
- Focus on student strengths (language competence) rather than weaknesses.
- Attention to learning styles, student background, and language level.
- Authenticity due to activities that exercise learning objectives, lead to course goals, and reflect tasks that are required for classroom and real-life functions.

Similar to an artist or architect's portfolio, a writing portfolio consists of a variety of writing samples produced by a student over an extended period of time. Often, in a

writing class, students compile their portfolio out of the best selection of work they have produced throughout their composition course. Under guidance of their writing instructor, students collect those writing samples that they feel best represent their writing ability. This collection of compositions, which may or may not include a series of drafts for each writing project, as well as self-reflective metacognitive statements in which students explain what they learned during the course, is then submitted to the portfolio evaluator for grading. The exact contents of a portfolio will vary depending on the writing program; however, leading researchers offer the following as characteristics of good portfolios. Hamp-Lyons (1994) offers the following nine points:

- *Collection*: more than a single writing sample
- *Range*: a variety of writing genres
- *Context*: strength through writers' expertise
- *Delayed Evaluation*: permits students to revise work
- *Selection*: students participate in choosing samples
- *Student Focus*: students take responsibility for portfolio success
- *Self-reflection*: includes metacognitive self evaluation
- *Growth*: evaluators can assess learning process
- *Development*: evaluators can observe the progression of a writing sample

Moya and O'Malley (1994) provide a comparable list that highlights many of the same points:

- *Comprehensive:* a variety (both breath and depth) of student work is included
- *Planned:* purpose, contents, schedule, and grading criteria are predetermined
- *Informative:* purpose and meaning are clear to students, evaluators, and others

- *Customized:* portfolio is adjusted to meet the specific needs of a program
- *Authentic:* samples reflect useful activities as part of course work

Both of these lists have several key ideas in common. These researchers emphasize that portfolios should include a variety of student work, meet an instructional and evaluative purpose, involve students in the selection of writing samples, and reflect authentic writing tasks.

Advantages of Portfolio Assessment

Portfolio assessment, as a form of performance testing, is valuable due to its ability to provide a more complete representation of learner ability. Additionally, writing portfolios have unique advantages in ESL composition courses. The benefits of portfolios include positive washback effect, consideration for the writing process, fairer testing conditions for ESL learners, and less exam preparation time for test creators.

As mentioned earlier, performance tests tend to produce greater positive washback effects than indirect forms of assessment. Portfolios, in particular, are agents for improved classroom instruction. This is because classroom writing activities become the basis by which learners are assessed at the end of the course. As such, there is no conflict between curriculum objectives and testing procedures. Teachers provide instruction and assign writing assignments that reflect the course objectives as designed by the administration. In turn, students select samples for their portfolio from among these writing activities and projects. Then, when they are graded based on their portfolios, they are assessed on their classroom writing assignments which were based on the curriculum objectives. It is a desirable harmony between instruction and assessment.

For the majority of the 20th century, writing instruction was primarily product-focused. This meant that students were taught to analyze and mimic literary works that possessed the qualities of what was defined as good writing. Little attention was given to the manner in which students arrived at the desired product; emphasis was placed on the qualities of the final product. In the 1970s, writing instruction shifted from a product- to a process-focused approach. Composition instructors believed that the method used to create good writing was the key to writing success. If students could be taught to employ an effective process when writing, then there would be greater chance that the final product would be of good quality. The grading of writing shifted from summative to formative evaluation.

Now, in the 21st century, writing program designers tend to use a combined product- and process-focused approach to composition. As a result, they require a form of assessment that addresses both product- and process-focused writing. Traditional performance-based writing tests, such as the timed essay, do not account for the process that writers undergo in order to produce a piece of writing. Timed essays truncate the amount of time that learners have to plan and write a composition thus creating an artificial time constraint on the writing situation. Additionally, timed essays typically only include one draft of the test taker's composition which does not reflect the process-focused approach of current writing curricula which includes revising and redrafting. Portfolio compositions, on the other hand, are developed over time using authentic processes involving prewriting, researching, drafting, and revising tasks. Portfolios also allow for product-focused attention in that teachers and students can include a variety of different writing genres in the portfolio as required by the assessment criteria.

In recent years, numerous studies have been performed in order to verify the perceived advantage of writing portfolios as a fairer and a more accurate form of assessment for ESL writers. Many of these studies, such as Coombe and Barlow (2004) have been qualitative and case study in nature. However, Song and August (2002) conducted a longitudinal quantitative study on writing portfolios at Kingsborough Community College of the City University of New York (CUNY). The administration at CUNY requires all students to pass the Writing Assessment Test (WAT) – a direct, but timed, test of writing ability. The researchers noticed that, on average, ESL students had far more difficulty passing the WAT than their native English-speaking counterparts. This bias against L2 learners led the researchers to compare two sections of ENG C2 (first semester ESL freshman English), both which required students to develop portfolios throughout their coursework. At the end of the semester, the control section was assessed using the WAT, whereas the experimental section was assessed based on their portfolios. The researchers found that twice as many students passed ENG C2 when assessed using portfolios as compared to those who were assessed using the WAT. The following semester both sets of passing ENG C2 students were placed in ENG 22 (the second semester freshman English course). The researchers found that portfolio-assessed students passed ENG 22 with equal ratios and grade distributions to the WAT-assessed group. It is worthy to note that portfolio assessment appears to be twice as effective in identifying the number of ESL students that are prepared for the next level of English writing instruction. Song and August's results suggest that portfolio assessment is a fairer and more accurate form of testing for ESL learners than traditional timed essays.

Finally, it is intuitive that a portfolio-based assessment is easier to create than other forms of writing assessment: there are no multiple choice questions to write or evaluate; there are no timed-essay writing prompts to create and rotate each semester; there is less concern about ensuring that test items are reflective of the broad range of teaching objectives. Portfolios are a natural result of classroom instruction. Although portfolios appear to be the easier route to writing assessment, they create several other time- and validity-related issues which will be discussed in the following section.

Problems with Portfolio Assessment

In their study of portfolios in EFL university settings, Coombe and Barlow (2004) found that one of the greatest challenges to successful portfolio implementation was the time involved in helping the students prepare sufficient metacognitive self-reflection for each portfolio entry. The researchers described two case studies wherein reflective portfolio assessment was implemented in university composition classes in the United Arab Emirates. Based on previous research that suggested metacognitive self-reflection is an essential part of effective portfolios, the course instructors had intended for their students to include reflective elements for each entry into their portfolios. The researchers were surprised to learn how much time and effort it took for students to comfortably and reliably write reflective statements about their portfolio entries. As a result, the number of entries for the reflective portfolio was reduced in order to meet the course assessment deadline. The researchers advised instructors and administrators to be aware of the potential time it may take to train students in the metacognitive self-reflection that is a recommended element of effective portfolio assessment.

Another time-related problem with writing portfolios relates to rater training. Because portfolio assessment is a performance-based form of testing, it requires a greater investment in post-exam evaluation than indirect testing. Many indirect test items (such as multiple-choice or fill-in-the-blank) can be quickly and objectively graded, whether by hand or even by computer. Portfolio assessment, as with most direct tests of writing, requires the use of human raters. Raters are unavoidably subjective in their judgments, even when provided with a detailed rating scale and criteria. Fortunately, research suggests that several effective rater-related elements (including rater training and pre-grading collaboration) can help increase reliability and consistency among human raters (Lumley, 2002). So although administrators may prefer portfolio assessment because it will save time in test creation, this form of assessment requires a large investment of time in proper rater training and exam grading.

Authenticity of student work is another concern that some administrators have with portfolios. Because portfolios usually consist of writing samples created for course work, there is greater potential for students to receive outside help (i.e., tutors, family, friends, etc.) and thus place in question the degree that their portfolio work is an accurate assessment of their own work. Proponents of portfolio assessment have suggested that a timed essay sample could be included in the portfolio as a benchmark to help graders assess the unaided writing ability of the student. Unfortunately, this solution brings with it many of the negative aspects of timed essays; this writing sample does not necessarily measure writing ability in everyday use. It may be complicated by test anxiety issues, limited planning and revision time, and lack of editing and consultation sources that writers frequently use when completing authentic writing tasks. If a timed essay sample is

used in portfolio assessment, it should be viewed as a measure of general writing fluency, and not as a comparison for true potential in common writing situations.

Some opponents to portfolio assessment also express concern that, in many cases, students do not receive a grade throughout the semester and have no indication of their ability before the grading of the portfolio at the end of the semester. Although this may be the case in some implementations of portfolio assessment, other programs encourage teachers to grade student essays throughout the semester, providing meaningful feedback so that student will not only have an indication of how well they are achieving course objectives, but so that they will also receive guidance on how to improve their writing so as to meet the objectives before the final grading of the portfolio. However, this raises yet another problem: the potential discrepancy between teacher grading and rater grading of the same work. A student could receive positive feedback from a classroom teacher only to be graded harshly by a portfolio rater who interprets the course objectives differently than the classroom teacher. This concern re-emphasizes the need for effective rater training and for a shared understanding of grading and course objectives among teachers and raters.

Another criticism of portfolio assessment, especially holistic grading, is that such a form of assessment does provide learners with enough specific feedback to help them improve. Holistic grading is a general evaluation of a writer's ability: high holistic scores do not inform examinees of areas of success any more than they inform of areas that need improvement. Proponents of this criticism should consider that final exams are primarily designed to be achievement tests and not sources of performance feedback; the responsibility for meaningful, formative feedback should remain with course instructors

and tutors. Although ideally an exam should be able to make evaluative decisions as well as provide useful feedback, in reality this duality is difficult to balance due to time constraints on exam raters. The more feedback raters are required to provide, the less time they will have to grade efficiently and reliably. Administrators may conclude that the purpose of final exams, such as the LATs at the ELC, is to help make a decision about examinee proficiency and not to provide detailed feedback.

In addition to all these issues, one concern is consistently raised in the literature related to writing portfolio assessment – the validity and reliability of raters' use of the rating scales and criteria (Arkoudis & O'Loughlin, 2004; Bachman, 2002a; Bachman, 2002b; Lumley, 2002; Lynch & Mason, 1995; Song & August, 2002). Schoonen (2005) admits that "it seems a price is paid for [the advantages of portfolio assessment]: the complexity and multi-faceted nature of performance-based testing introduce multiple sources of error" (p. 2). It is because of this concern that a writing test requires a validation study in order for test administrators to place confidence in the assessment results.

A Validation Study

Nature of Test Validity

Test validity is the degree to which an exam (or alternative form of assessment) can be relied upon to give meaningful results and then the degree to which those results are appropriately applied to further situations. Although in the past some researchers have described validity in terms of different types of validity, the common view today is that of a unified validity (Messick, 1992; Weir, 2005). Unified validity consists of several types of validity-related evidence that together help establish the degree to which an

instrument is valid. No single source of validity can make this claim, nor is any one source considered to be superior to another. Instead, all types of validity-related evidence complement one another in a researcher's efforts to measure validity.

Messick (1992) describes validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment" (p. 1487). A test with a high degree of validity is one which measures what it is designed to measure, and whose results are interpreted and applied to meaningful situations about the test takers and their ability to succeed in appropriate situations.

For example, many motor vehicle departments require motorists to pass a written test of driving knowledge. Although such tests can be said to portray an accurate reflection of the test taker's passive knowledge about driving, a community would be hesitant to issue driver's licenses based exclusively on the written test. Instead, most governments require a performance exam – a driver's road test – in which the potential license holder is required to demonstrate driving knowledge in a real-world, hands-on situation while being assessed by a rater.

Having passed this form of assessment, test takers will then be issued a license and be declared fit to drive an automobile; the exam is said to be an accurate reflection of both the examinees' current ability and their potential to perform adequately in future driving situations. Those who do not score high enough on the road test are deemed to be unfit and ill-prepared to drive; these candidates will not be issued a driver's license until they achieve a passing score on the rated road test. The driving exam, however, should not be used as an accurate indicator of the licensees' ability to fly airplanes or conduct

sea craft. The exam is valid so far as it measures what it is designed to measure, and to the degree that its results are applied to relevant situations.

Need for a Validation Study

The purpose of a validation study is to gather various types of evidence to help assess to what degree an instrument is functioning as expected, and also to what degree the results are being properly interpreted and used. In order to make this evaluation, a researcher collects evidence from various aspects of the testing instrument and its applied context. It is not sufficient to make a claim of validity based on one single type of evidence; rather, a researcher will seek out numerous forms of support to establish to what degree – and not a definitive statement that – a test is valid. By necessity, a well-grounded validation study incorporates a myriad of research questions in order to make a statement about an instrument's degree of validity. Validity studies help administrators and educators place confidence in the scores and applications of tests.

In the driver's license example, an examinee may be required to provide more than one source of predictive evidence of good driving ability. The motor vehicle department may require an eye-exam, may ask for a doctor's recommendation of good health, and may conduct a criminal background check on the potential licensee in order to gather information regarding the test taker's value as a healthy, responsible, and conscientious citizen. These various facets about the examinee's ability provide a more complete indication of the examinee's potential to meet the standard of a good driver. In the same way, a good researcher requires multiple sources of evidence to make a more substantive claim about the validity of an exam.

Bachman (2002) stresses the importance of both qualitative and quantitative data when conducting an evaluation of a task-based test. He suggests that researchers “use a variety of procedures to collect information about test performance, along with a variety of analytic approaches, both quantitative and qualitative, to tease out and describe in rich detail the interactions among specific tasks and specific test-takers” (p. 471). A researcher can make a stronger case for validity when both qualitative and quantitative evidence are collected and analyzed in relation to one another.

The driver’s license metaphor can also be used to demonstrate how qualitative and quantitative evidence are essential to an effective validation study. If an independent government agency was assigned to assess the validity of the driving exam, the investigators would gather information from numerous sources in order to evaluate the degree to which the motor vehicle department’s driving exam procedures were successful in differentiating between good and bad drivers. The researchers from the independent agency might randomly choose a selection of people who had recently been issued licenses by the motor vehicle department and question them about their driver’s exam experience. Evidence regarding the test takers’ potential as responsible drivers would be gathered from numerous sources, such as another written test, a second road performance test, a criminal background check, interviews with examinees regarding their driving practices, interviews with road test raters regarding their use of the rating scale and criteria, rater consistency and inter-rater reliability studies, and any accident or insurance reports related to the licensees. If negative or conflicting results were gathered in one or more of the above areas, doubt could be cast about the degree of validity of the driving test. Reliance on only one type of evidence would weaken the agency’s evaluation of the

exam; multiple sources would provide a more complete assessment of the effectiveness of the driving exam. The researchers might suggest some areas for improvement in order to increase the driving test validation. Their objective, and thus the objective of any validation study, is to gather sufficient evidence – both qualitative and quantitative – to make a judgment regarding the effectiveness of the exam and the appropriateness of its use.

Types of Validity-related Evidence

The types of evidence on which a validity study relies can vary. Validity evidence can be categorized into many different categories (see Bachman 2002a, Brown 2000, Cumming & Berwick 1996, Hughes 2003, Kane 2001, Kumar 1999, and Park 2004). The purpose of different categorizations is to help researchers understand and emphasize multiple ways in which validity should be considered. For example, a program administrators should consider aspects of construct validity (is the test measuring the desired trait), content validity (does the test measure an adequate portion of the learning domain), and criterion validity (how does the test compare to other indicators of the desired trait). Messick (1992) proposes that researcher approach a validity study with the aim to investigate a unified validity. This interpretation of validity helps emphasize that no particular aspect of validity is more important than another; in order for an administrator to gain confidence in the validity of a testing situation, multiple aspects of unified validity much be confirmed.

To attempt to account for all possible interpretations and categorizations of validity and reliability in a single study is a daunting task, and most likely an unrealistic one. This presents a researcher with an initial dilemma: if a good validity study is one

that accounts for a wide variety of validity-related evidence, and yet to undertake such a task is unmanageable, what should a researcher do? A helpful response would be to select one aspect of validity for an initial study, to gather various sources of evidence related to that one aspect, to analyze and interpret them, and then, in further studies, to move onto other aspects of test validity. There is simply too much to do at once; however, conscientious researchers will employ a recursive process as they work to evaluate and improve the effectiveness of an exam.

This study focuses on scoring-related validity, a term which Weir (2005) uses where others use reliability. More specifically, this study describes processes for evaluating the scoring-related validity/reliability of a writing exam, which includes an evaluation of levels, classes, examinees, raters, writing criteria, and the rating scale.

Scoring-related Validity

Reliability as Scoring-related Validity

In any well informed discussion on test validity, the researcher will include an explanation of reliability, the degree to which something is produced consistently. For many researchers it is two separate, though related, investigations: validity and reliability. However, Weir (2005) takes a different approach. When reliability refers to a test's score, Weir views this as an extension of validity-related evidence and refers to it as *scoring-related validity*. He states that the traditional separation of reliability from validity is "unhelpful and [that] reliability would be better regarded as one form of validity evidence" (p. 14). Weir's approach sees score reliability as yet another area in which validity can, and should, be established. When conducting a validation study, researchers

should address issues of scoring-related validity in addition to the more traditional aspects of context-, content-, and criterion-related validity.

This study follows Weir's classification; hence, when reference is made to scoring-related validity, the reader should keep in mind that it is a term that is frequently referred to as "score reliability" by other researchers. The majority of literature on the subject still uses the term "reliability" rather than "scoring-related validity;" however, in this study both terms are used interchangeably, the former employed when discussing the work of others, and the latter used when referring to the current study.

Scoring-related validity is more than reliability or consistency; it involves the process, the differentiation, the meaning, the interpretation, and the usefulness of exam scores. Reliability tends to be concerned simply with the degree to which the same testing situation (examinee, rater, items) will produce identical results on retest. Scoring-related validity includes this concept but goes beyond this definition. A test that has a high degree of scoring-related validity is one that has a consistent rating process and is also meaningful. The test helps evaluators interpret and apply scores to appropriate situations. The test has a positive impact of teaching and learning. So, although scoring-related validity may be confused with terms such as reliability or consistency, it is in fact a much more complex concept. Scoring-related validity is the reliability and applicability of an exam score.

In order to provide a better understanding in regards to the importance of scoring-related validity, this section will describe numerous studies that investigate this issue in regards to performance-based language tests. Many-Facet Rasch Model analysis studies, generalizability studies, and qualitative studies will all be discussed.

Many-Facet Rasch Model Studies

Traditionally, inquiry into test validity has been done through classical test theory (CTT). However, the 1960s gave rise to a new approach to measurement theory: item response theory (IRT). The need for item response theory arose when researchers wanted a model that would account for the relative difficulty of test items when assigning a score to an examinee. For example, if an examinee of lower ability correctly answered a more difficult item, how can a test reward that student? Likewise, if an examinee of remarkable ability incorrectly answered a relatively easy question, should the test penalize the student more than if he missed a relatively difficult question? CTT cannot account for the difference in difficulty of test items. Instead, test-makers must refine the test over time and possibly place a greater weight on more difficult questions.

IRT, on the other hand, takes a dynamic approach to test evaluation. It weights examinee performance based on the responses of other examinees such that examinees who respond favorably to more difficult items are given a higher trait measure than those who only respond favorably to easier items. This type of analysis is more sensitive to the interactions between examinees, items, and other possible exam facets. Emberton and Reise (2000) briefly state the difference between CTT and IRT. They explain, “IRT is based on fundamentally different principles than CTT. That is, IRT is a model-based measurement that controls various confounding factors in score comparisons by a more complete parameterization of the measurement situation” (p. 8). When a testing situation requires evaluators to account for the confounded interaction of exam factors, then IRT is a more appropriate measurement theory than CTT.

One model within IRT is the 1-PL (one parameter) Rasch model. Initially created by Danish mathematician Georg Rasch in the 1960s and subsequently developed by European and American educational and psychometric researchers, the Rasch model allows test evaluators to measure the combined interactions of multiple exam facets while taking into account the difficulty (or in some cases ability, severity, etc.) of each item within each facet. In other words, the Rasch model allows evaluators to account for the severity of a rater, the difficulty of a test item, and the ability of an examinee when tallying exam scores. For example, examinees' scores are adjusted to account for more difficult tasks or more severe raters. Likewise, the difficulty of a test item is more fully understood when the relative ability of respondents is more clearly understood. Rasch modeling allows researchers to account for these interactions and adjust accordingly.

The Rasch model accomplishes this by first assigning preliminary trait scores to each item in each facet. For example, each student is given an initial ability estimate along an ability axis. Then, the positioning of the student along the axis is adjusted as recursive computations are done to account for the relative trait levels of the intervening facets, such as test item difficulty, rater severity, etc. The model continues to adjust measurements (expressed using logarithm scale units called logits) until the desired degree of specificity is achieved. In this way, Rasch modeling accounts for variation in facet items and gives a more accurate sense of trait levels for examinees, test items, raters, etc. Sometimes, however, the model cannot account for all the variation in one (or more) items. These items act in inconsistent ways, are assigned a higher infit value than items whose variation can easily be accounted for, and are labeled misfit. Examples of misfit items include low ability examinees who inconsistently score well on some high

level test items while missing lower level ones, test items that do not follow a consistent pattern in distinguishing between higher and lower level examinees, and raters who are inconsistently severe or lenient in their grading of examinees or test items.

Today, an increasingly popular research tool for assessing test validity is Rasch modeling computer software. Two of the most popular packages, Winsteps and its many-facet counterpart FACETS, allow a researcher to quantitatively analyze a testing situation to investigate the presence of anomalous data that suggest that a test is not functioning in a valid manner. Validity, in regards to Many-Facet Rasch Model (MFRM) analysis, can be assessed in two primary ways: the degree to which items in a facet vary (examinee variance is desirable but other variance usually is not), and the presence (or lack) of misfit items. In a sense, a test could be defined as valid if MFRM software shows that items vary accordingly and that there are no misfit (i.e. anomalous) data in an analysis of test score results.

This is the approach taken by Park (2004) in an investigation of scoring-related writing exam validity. Writing tests scored by human raters are subject to scoring-related scrutiny. Administrators and other stake-holders want to establish the validity of the exam scores given the potential for error, variability, and subjectivity in test scores. Park's study demonstrates how MFRM software (in this case FACETS) can be used to evaluate the effectiveness of a writing exam. Rather than attempt to account for various types of validity in his study, Park employs a "restricted definition of validity…one that is common in Rasch analysis: if Rasch analysis shows little misfit, there is evidence for the construct validity of this measurement procedure" (p. 3). This interpretation of validity allows a researcher to make an initial judgment of test validity based simply on

software analysis of sample exam scoring data: in this case, examinees, raters, performance criteria, and the rating scale.

Park's MFRM analysis showed little misfit data. Some anomalous data was flagged by the software, including misfitting examinees and biased criteria. However, the low level of misfit data in relation to the test as a whole not only provides test administrators with a quantifiable measure of test validity, but the analysis also can help administrators identify aspects of the exam that need to be improved, such as rater training, criteria clarification, or scale adjustment. MFRM is an effective quantitative tool for measuring test validity and is also useful in spotting areas for improvement.

Generalizability Studies

Researchers may conduct a generalizability study as another method of quantitative inquiry used to help establish the degree of scoring-related validity. Brennan (1992) describes a generalizability study as research that attempts to estimate optimizations based on multi-variant facets in a process. When applied to the rating of a writing exam, for example, a generalizability study might be designed to predict the most efficient number of test task items, writing portfolio samples, and raters required to produce consistently valid results. Researchers in a generalizability study might conclude, based on a series of multi-variant statistics, that administration can expect to get the most valid test results by using three task items and four portfolio samples rated by two different raters; using more raters could be a waste of effort since two raters are found to be as effective as three. Likewise, the researchers might suggest that using fewer than 4 portfolio samples drastically lowers the degree of validity below the desired level. A good generalizability study (G-study) will provide the most efficient combination or

balance of variables involved in a testing process in order to save time and money while maintaining an acceptable degree of test validity.

G-study software, such as GENOVA, helps provide high-level assessments of exam functioning. For example, G-study software can predict the reliability coefficient of a test if a variety of facets are reduced/increased, such as the number of student writing samples or the number of raters used in grading the exam. G-studies can also help account for error in a testing situation, helping researchers identify whether variance comes from persons (desired variance among examinees) or from unwanted sources such as raters, writing samples, rating occasions, or interactions between these and other facets.

Schoonen (2005) conducted a generalizability study of a performance-based writing test administered to eighty-nine grade six students. Several variables were considered in the study including task (describing, narrating, etc.), number of raters, number of student writing samples, scoring procedure (analytic versus holistic), and rating criteria (content and organization versus language use). Schoonen found that based on four writing samples from each student, there is little use in hiring more than two or three raters; four raters did not significantly increase the reliability. However, this was dependent on rating criteria and scoring procedure. When students were rated based on language use using holistic scoring, a minimum of two raters were able to provide sufficient reliability. At the same time, when content and organization was rated using analytical scoring, more raters and writing samples were needed to achieve an acceptable degree of scoring-related validity (based on a generalizability score of 0.80). In conclusion, Schoonen suggests that portfolio assessment may be one of the most reliable

forms of performance-based writing assessment since it is the number of writing samples, and not the number of raters, that appears to have the greatest influence on scoring-related validity. Performance assessments that include only one writing sample, such as timed essay exams, are not as reliable as multiple sample writing exams even when several raters are used. Additionally, Schoonen recommends that test developers further investigate rating criteria in order to improve test validity.

Several other researchers have found that a combined MRFM analysis and G-study can cooperatively shed light on test efficiency (Sudweeks, Reeve, & Bradshaw 2005; Bachman, Lynch, & Mason 1995; Lynch & McNamara 1998). These researchers argue that a combined use of G-study and MFRM software offer stake holders both an evaluative as well as a predictive appraisal of writing assessment situations.

A G-study gives a more macroscopic view of test reliability among the facets in question, whereas MFRM works to identify specific elements (examinees, criteria, raters, etc.) that are not functioning in a reliable manner. Although it is advantageous to use both analyses when conducting a validity study, most generalizability theory software requires a fully-crossed design (i.e. all raters rating all writing samples of all students' exams). The cost and time involved in a fully-crossed design is difficult to achieve, especially in a study that accounts for more than 200 students and over 800 combined writing samples such as the ELC's writing portfolio LAT. MFRM software, on the other hand, does not require a fully-crossed design and can be a more economical choice. Ideally, both a G-study and a MFRM analysis should be done, but in the event that a G-study is not feasible, MFRM alone can still give researchers a richly meaningful view of test validity both macro- and microscopically.

Qualitative Studies

Numerous researchers have qualitatively investigated the validity of performance-based testing including two recent works by Cumming (2001) and Huot (2002). However, Lumley (2002) insists that still more research needs to be done on performance-based testing before researchers and administrators can feel confident in the medium's degree of validity. Lumley is particularly concerned how the choice of rating criteria will influence test validity. He speculates that if rating criteria are effectively chosen and applied, high scoring-related validity will occur, while raters' inconsistent use of and misunderstanding regarding the rating scheme will drastically lower validity. As a result, Lumley investigated the effect of assessment criteria on raters' scoring decisions.

Lumley's (2002) study involved the use of Think Aloud Protocols (TAPs). A qualitative research tool, TAPs require human subjects to complete the task under investigation while voicing aloud their thought processes. TAPs are one of many verbal reporting methods described by Gass and Mackey (2000). They describe the technique saying, "Despite different terminology, verbal reporting can be seen as gathering data by asking individuals to vocalize what is going through their minds as they are solving a problem or performing a task. Verbal reporting allows researchers to observe how individuals may be similar or different in their approach to problems" (p. 13). TAPs can be an effective qualitative research tool that enables test evaluators to investigate how their subjects internally approach the task under investigation.

There are several benefits as well as challenges to TAPs. TAPs are effective because they are an "online" technique for gathering subject data (meaning that the data gathering happens in real time), there is little-to-no time between the action and the

reporting of the action. This results in less intervening factors; the subject is not influenced by situations that take place between the time that the task was completed and the time that the task was reported. TAPs are also used because not only do they help reveal information that is unobservable, but also because they reveal processes and not just discrete information. However, TAPs have some possible weaknesses. As with most human subject observation techniques, there is the threat of halo effect: subjects may adjust their process as a result of being observed. They may wish to please the researcher by producing what is believed to be the desired process, or they may become nervous and perform the task in a manner inconsistent to their normal process. Also, there is threat of inconsistency in the reporting of TAPs. Unless a consistent method for TAPs observation and interpretation is employed, data may be confused and misused.

Protocol analysis researchers Ericsson and Simon (1984), along with Gass and Mackey (2000), admit that verbal reporting faces these challenges. Together, they offer suggestions for mitigating these concerns by following a few guidelines. First, whenever possible, subjects should be chosen that feel comfortable being observed. They should be instructed to perform the task as they normally would except for voicing aloud their thought processes as they do so. Next, subjects should not be informed of the specific aims of the research or any hypotheses; instead they should be told that the researchers are just looking to document the subjects' natural processes. Finally, ideally one observer should conduct all the TAPs. This will lower the chance that data will be reported differently from one observation to the next. This same observer should be on hand to interpret the data. Reliability and consistency in recordings can also be improved through

the use of audio/video equipment and even the presence of a second observer who also collects the same data and then verifies observations with the primary observer.

Using TAPs, Lumley (2002) analyzed how raters in a large-scale writing assessment employed the rating criteria as they assigned scores to test takers' compositions. He found that despite rater training and explicit wording in the rating scheme, raters frequently hesitated and struggled to assign scores to test takers' work. Moreover, Lumley observed much variation in the manner in which raters approached a text and rationalized the score they assigned. Surprisingly, rather than use the rating criteria to assign a score, raters tended to make an intuitive judgment about a composition and then attempted to justify that decision based on the rating scheme. One reason for this may be that the scheme lacked criteria that the raters felt were important, such as length or quality of ideas. At other times raters were frustrated by the lack of priority among criteria data; they felt some factors were more important than others, but the rating scheme did not allow for this.

In order to overcome this inconsistency, Lumley does not necessarily suggest that rating schemes be improved to include criteria that raters value but do not currently exist in the scheme. Instead, he proposes rater training that encourages raters to view the rating scheme as a guiding tool rather than a set of binding rules. He also encourages administrators and evaluators to accept that a "true" rating scheme is essentially unattainable; rather, rating schemes should be viewed as "a set of negotiated principles that the raters use as a basis for reliable action, rather than a valid description of language performance" (p. 268). Despite Lumley's conclusion that no "true" rating scale can ever be created, he does not altogether abandon the development of rating scales and criteria.

He suggests that evaluators do their best to design valid content- and construct-related rating schemes that can then be used to help raters achieve scoring-related validity.

The development of rating schemes is the topic of a study performed by Turner (2000). She recorded and analyzed the discussion among a group of teacher-raters who reviewed student essay responses to an ESL writing exam in order to identify the qualities of good writing. Turner emphasizes the need for a shared understanding of rating criteria between raters; she also states that a rating scheme be based on actual student writing. In her study, she details how a group of ESL teachers and administrators in Quebec read student essays and then categorized them into various piles according to perceived ability. Then, the essays in each pile were analyzed and the common traits of writing in each category were listed. This enabled the group to generate a list of writing traits that would then be used to rate future exams.

By observing the group negotiate the criteria of good writing, Turner was able to report not only how teacher-raters can disagree on rating criteria, but she was also able to conclude that rater calibration and discussion before rating is essential if a writing exam hopes to achieve reliability. She states that,

> The fact that teachers developed the scales brings with it discourse stances, beliefs, and understandings of the TESL curriculum that are very specific to the context. This may have a positive impact on such factors as inter-rater reliability when the scale is used within its intended area. It is to be noted that much discussion is involved in the scale making process because it involves working with actual student performances and coming to a consensus, … In other words, teachers need to work out differences. (p. 576)

Whether in the development of a rating scheme, or in the use of it, insight can be gained from observing raters discuss and voice their preferences and interpretations of writing criteria.

Arkoudis and O'Loughlin (2004) use discussion groups and interviews with teachers to investigate the validity of an ESL writing exam in Australia. Teachers in this study expressed concern and frustration over their roles as writing evaluators. Their concerns included difficulty interpreting, then applying, the evaluation guidelines to actual student writing. Some teachers felt it was easier to assess lower level students, where criteria were fairly simple, than upper level students where the criteria was more confounded. They also were troubled over the responsibility they felt to achieve intra- and inter-rater reliability.

Through these interviews and discussions, the researchers were able to help the teachers create a list of concerns and proposed solutions regarding the implementation of the writing exam. This list was then shared with administrators. Arkoudis and O'Loughlin (2004) conclude that the process of consulting teachers and then sharing their concerns with administration has several possible benefits. First, it helps teachers come to a common understanding of the challenges they face as raters. This helps them recognize that they are not alone in their concerns as evaluators. Second, the discussions included negotiation of writing criteria and other rating issues that may help improve intra- and inter-rater reliability. Additionally, by sharing their concerns with administration, teachers may feel an increased sense of cooperation towards administration. Finally, it is hopeful that the administration will thoughtfully consider the teachers' concerns and pursue changes to the testing context that could help improve the rating process.

This section has summarized qualitative studies that reveal how researchers can involve teacher-raters in a writing exam validation process. Whether it be through TAPs, discussion groups, or interviews, teacher involvement in qualitative research brings valuable insight to a validity study that cannot be gathered through quantitative analysis alone.

Conclusion

This chapter has provided a definition of the writing construct and offered reasons for the testing of writing. Numerous methods of testing have been offered with a discussion regarding the advantages and disadvantages of each. In particular, an alternative form of direct, performance-based assessment was introduced – the writing portfolio. Benefits of portfolio assessment were detailed including its positive washback effect, its authentic reflection of external writing skills, and its ability to more accurately and fairly assess ESL learners. Following this, problems with portfolio assessment were introduced, including the issue of test validity. The concept of validity was defined as a multifaceted construct. Various aspects of validity were shown with an in-depth discussion regarding scoring-related evidence of validity as it applies to raters and their use of rating scales and criteria.

The studies discussed in this chapter illustrate the need for research that combines both qualitative and quantitative data in an effort to validate a writing portfolio exam. Furthermore, current research into portfolio exam validation has only focused on single courses; no study demonstrates how neither qualitative nor quantitative research methods can be applied to a validity study of a multi-level ESL writing portfolio program. As such, this literature review has provided a basis for the research design of the current

study – a combined qualitative and quantitative validation inquiry into the scoring-related evidence of the multi-level writing Level Achievement Test (LAT) at the English Language Center (ELC) of Brigham Young University (BYU).

CHAPTER THREE

Research Design

Chapter Two provided a theoretical basis and practical models for a validation study of the rating scale and criteria of the writing Level Achievement Test (LAT) at the English Language Center (ELC) of Brigham Young University (BYU). Chapter Three applies the literature to a plan for conducting an investigation into the validity of the writing LAT's rating scale and criteria. First, a description of the ELC's writing LAT is provided, followed by an account of the subjects involved in the study. Then the study's four primary validation analyses are described: a FACETS Many-Facet Rasch Model analysis, a rater survey, rater Think Aloud Protocols, and rater follow-up interviews. The following research questions guide the explanation of this study's research design:

Quantitative evidence:

1. How valid are the writing LAT scores based on a Many-Facet Rasch Model analysis? This question is further separated into the following questions.
    a. How well do the LATs distinguish among levels, classes, and examinees?
    b. How severe are the raters in relation to one another?
    c. How consistent are the raters?
    d. How difficult are the writing criteria in relation to one another?
    e. How well is the rating scale used?

Qualitative evidence:

2. What is the degree of rater agreement on the priority of rating criteria among and between levels?

3. How do raters apply the rating criteria to determine a LAT score? (i.e. prioritizing some criteria more than others, and valuing some criteria more in higher levels than in lower levels, etc.)
4. How do raters use portfolio samples to negotiate a LAT score? (i.e. holistically based on all scores, or an average based on individually determined scores for each sample, etc.)

Description of the ELC's Writing LAT

The writing LAT at the ELC is administered at the end of each semester (April, August, and December). During the last week of class, students, with the help of their teachers, select two writing samples from their collection of multi-draft essays created as a requirement for their writing class course objectives (the number of assigned essays from students can select portfolio samples varies slightly from class to class and level to level depending on the teacher, but is usually 5 samples for levels 1-3 and 3 or 4 samples for levels 4 and 5). Also, during the final week of instruction, students write a self-reflective ("metacognitive") composition describing their development as a writer during the time that they wrote one of the chosen multi-draft essays. Finally, students write a timed 30-minute essay on an assigned topic during the semester's final exam days following the last week of instruction. This timed essay, similar to the TOEFL (Test of English as a Foreign Language) timed essay, is written in the computer lab (either on computer or by hand) during exam week. Upon completion of the timed essay, all four writing samples are collected into each student's portfolio and prepared for grading by the writing program raters.

In summary, the LAT is based on portfolio assessment and each student submits a portfolio that contains four samples of their writing ability:

1) One multi-draft essay;

2) A second multi-draft essay;

3) A metacognitive essay;

4) A 30-minute timed essay.

Once all the portfolios have been submitted for grading, the ELC writing program coordinator meets with all the writing raters for a rating calibration meeting. Raters are usually selected from the current semester's writing teachers as well as any other teachers as needed. Teacher-raters are required to participate in exam rating as part of their teaching contract; teachers are assigned to rate either speaking or writing LATs (grammar, listening, and reading LATs are all graded by computer). In this rater training session, the writing coordinator and the teacher-raters review rating procedures. Then in smaller, level-specific groups, raters discuss rating criteria and review benchmark essays until they feel they have a common understanding of the rating scheme and how it applies to the essays that they will read.

Following the rating collaboration meeting, all student portfolios are double-rated: each portfolio is first graded by one rater, and then by a second rater with the first score remaining blind until the second rating is complete. Readers assign a holistic score covering all four portfolio samples based on a set of criteria: topics, content, organization, vocabulary, grammar, editing, and the writing process. According to the holistic criteria, portfolios are assigned a score according to the rating scale. Portfolios that are found guilty of plagiarism or are missing samples/drafts are not assigned a score and are given

an immediate failing grade. This study is only concerned with individual ratings and the difference between them, but not with the final portfolio grade assigned to students (an average based on the two ratings). As such, the system of final percentage score averaging will not be explicated. At the end of the rating process all ratings and final scores are recorded and rater feedback sheets are distributed to students informing them of their portfolio score.

Description of Examinees

Subjects for this study are adult ESL students at the ELC in Provo, Utah. The ELC is an intensive English language program (IEP) operated by the Continuing Education and the Linguistics and English Language departments of BYU. The ELC provides English language instruction to adults who wish to improve their English for academic, vocational, social, or self-enrichment purposes. Data for this study will come from the LAT administered during the Fall 2005 semester. In total, 251 student portfolios were graded by raters for the Fall 2005 writing LAT.

Subject proficiency levels range from high beginning to low advanced. Students at the ELC are assigned to a proficiency level (1-5) based on placement and diagnostic exams administered at the beginning of each semester. The beginning-of-semester exams, as well as the LATs, are given in five areas: grammar, listening, speaking, reading, and writing. The number of students in each level is shown in Table 3.1. Students are both male and female and range in age from 18 years on. They come from more than ten different native language backgrounds and a wide variety of home nations.

Table 3.1

Examinees per Level

Level	Examinees (Count)
1	26
2	42
3	79
4	70
5	34
Total	251

Table 3.2

Raters per Level

Rater ID	Gender	Levels Rated	Number of Portfolio Groups Rated
R11	Male	1, 2, 3	3
R12	Female	1, 2, 3	3
R13	Male	2	2
R14	Female	3	2
R15	Male	1, 2, 3, 4, 5*	2
R16	Female	4	2
R17	Female	4	2
R18	Female	4	1
R19	Female	4, 5	2
R20	Male	5	1
R21	Female	5	2

*R15 served as the triple rater for all levels when needed

Description of Raters

All 11 raters are ELC teachers who were required to rate LATs as part of their teaching contract. Nine raters were current writing teachers for the Fall 2005 semester. The remaining two raters were reading teachers who had taught and rated writing at the ELC in previous semesters. Each rater was assigned to rate one set of portfolios for each class taught that semester. Raters were selected by the ELC writing coordinator based on her impressions and experiences regarding their ability to rate accurately and effectively. Raters were chosen from teachers of all five ELC levels and included both men and

women (see Table 3.2). All raters received rater training in the rating calibration meeting the week before the rating process began.

Description of Rating Process

Once students have completed their 30-minute essay and compiled their portfolios, the writing coordinator collects all student portfolios and divides them into portfolio groups for each level, usually with 10-15 portfolios per portfolio group. Each group contains a mixture of portfolios from each class in a given level. There are as many portfolio groups as there are classes taught by each of the raters. Because the raters for this semester taught a combined total of 21 classes, the portfolios were divided into 21 groups. The writing coordinator pre-assigns raters to groups such that each portfolio will be rated by two different raters. Double rating is mixed such that there is an overlap among raters. For example, Rater 11 graded one group of Level 3 portfolios. That same group was doubled-rated by Rater 12, who also graded a group that was doubled-rated by Rater 14, and so on. Maximum overlap among raters was achieved whenever possible.

In the morning of the first day of rating, the writing coordinator distributes the portfolio groups and rating sheets to raters (see Appendix B). Raters have until that evening to rate their portfolio groups and then return them to the writing coordinator (actual rating time typically takes between one to two hours per portfolio group). Once the first rating is complete, the writing coordinator records scores and prepares the portfolios groups for the second rating. The following morning the process is repeated; raters receive a new batch of portfolio groups and return them to the writing coordinator by the evening of the second day. The writing coordinator then records all second ratings. If there is a discrepancy of more than one scale point between the two ratings for a given

portfolio, then a third rater provides an additional rating (triple ratings are required for less than one percent of all portfolios). In either case – double or triple rating – the ratings are averaged and the portfolio receives a final score. Figure 3.1 provides a visual representation of the rating procedure.

Exam Day	**Rating Day 1**	**Rating Day 2**
Students write a 30min essay exam	Raters pick up their first batch of portfolios	Raters pick up their second batch of portfolios
30 min essays are added to students' writing portfolios	Rating	Rating
Student portfolios are submitted to the writing coordinator	Raters submit their first batch of ratings	Raters submit their second batch of ratings
Writing coordinator prepares portfolio groups for rating	Writing coordinator records first batch of ratings	Writing coordinator records second batch of ratings
		Triple ratings (if any) are conducted and recorded

Figure 3.1 Writing LAT Rating Schedule

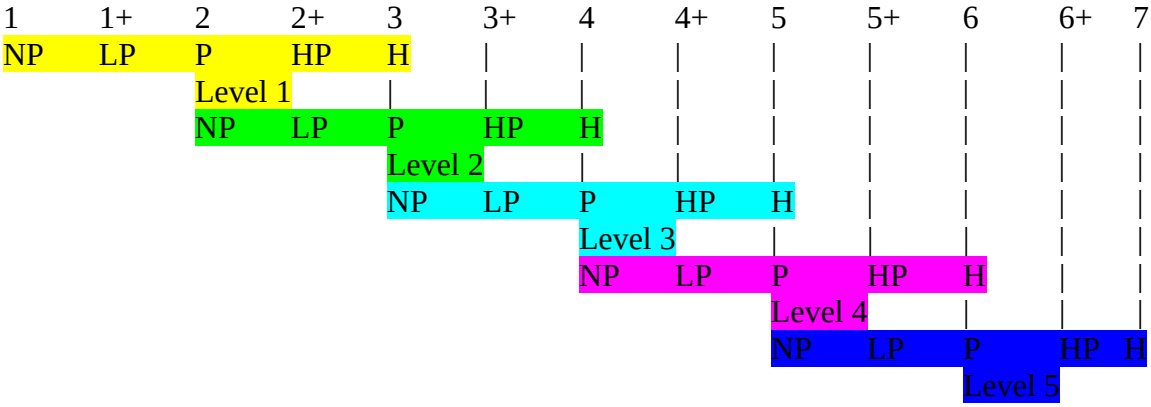
Description of Rating Scale

The writing LAT uses a 13-point continuous rating scale based on ELC proficiency levels (1-5) and two theoretical graduation levels (6 and 7). The scale (as shown in Figure 3.2) also has midpoints between each level (i.e. 1+, 2+, 3+, etc.). The points on the rating scale are designed to lend intuitive meaning to LAT scores. For example, a rating of 1 indicates that the student is writing at an ability of someone who is ready to begin the Level 1 writing class. A rating of 1+ indicates a student whose writing is higher than Level 1 beginner, but it not yet prepared to begin Level 2. Raters for a given level typically concern themselves with five points on the scale: the current level of the student, the midpoint between the current level the next level, the next level, the midpoint between the next level and the consecutive level, and the consecutive level. In

Fall 2005 Writing LAT Rating Scale

Continuous Scale:

- 13 points on the scale
- Raters generally use 5 points per proficiency level
- 5 points per scale are based on the previous scale's rating categories: NP (No Pass), LP (Low Pass), P (Pass), HP (High Pass), H (Honors).



What do the scores mean?

Score	Level 1	Level 2	Level 3	Level 4	Level 5
1	Needs to repeat Level 1				
1+	Will struggle in Level 2				
2	Ready for Level 2	Needs to repeat Level 2			
2+	Will do very well in Level 2	Will struggle in Level 3			
3	Possibly read for Level 3	Ready for Level 3	Needs to repeat Level 3		
3+		Will do very well in Level 3	Will struggle in Level 4		
4		Possibly read for Level 4	Ready for Level 4	Needs to repeat Level 4	
4+			Will do very well in Level 4	Will struggle in Level 5	
5			Possibly read for Level 5	Ready for Level 5	Needs to repeat Level 1
5+				Will do very well in Level 5	Will struggle in Level 5
6				Possibly read for BYU coursework	Possibly ready for BYU coursework
6+					Possibly will do very well in BYU coursework
7					Possibly will do extremely well in BYU coursework

Figure 3.2 Fall 2005 Rating Scale

other words, a rater grading portfolios of Level 3 students would normally use five points of the continuous scale: 3, 3+, 4, 4+, and 5. It is expected that a Level 3 student will not produce work below a rating of 3 nor above a rating of 5.

Quantitative Analysis

Many-Facet Rasch Analysis

The quantitative analysis for this study involved the use of a Many-Facet Rasch Analysis using FACETS item response theory modeling software. Many-Facet Rasch Modeling (MFRM) uses a 1-PL (one parameter) IRT (item response theory) model which allows a researcher to analyze the combined interactions between multiple facets. For the purposes of this study, the facets of interest were levels, classes, examinees, raters, and rating criteria. All the data from the LAT ratings (examinee level, examinee class, examinee ID, rater ID, and ratings for the overall and criteria scores) were collected into a single MS Excel document.

Once the data had been arranged in proper FACETS format, the Excel worksheet was exported as a .txt file which was then analyzed by FACETS software. Based on command file specifications, the software generated reports for every aspect of interest for this study. Specifically, the analyses of interest were logit scales for all facets, infit/outfit statistics for all facets, and category response curves for the rating scale criteria.

Qualitative Analysis

In addition to gathering quantitative evidence, this study also collected qualitative information regarding the validity of the writing LAT. The qualitative analysis used data from three tools: rater surveys, rater Think Aloud Protocols (TAPs), and rater interviews.

Rater Survey

The survey (see Appendix B) asks raters to rank the 14 rating criteria according to the importance they placed on each criterion when rating portfolios. These 14 items (topic difficulty, interesting content, length of papers, depth of topic, organization and order, depth/variety of grammar usage, accuracy of grammar usage, vocabulary, spelling, formatting, punctuation, writing process and drafts, 30 minute writing sample, and Metacognitive essay) were drawn from the criteria listed on the rater feedback sheets that are used to guide raters through the scoring process (see Appendix A). The feedback sheets contain only nine items; some criteria were expanded to create 14 items (i.e. *grammar* was split into *grammar accuracy* and *grammar depth/variety*). There is additional space on the survey for raters to indicate any additional criteria that they used to rate portfolios that was not included in the standard 14 items. The survey also asks raters to indicate the level that they rated; raters who rated more than one level were asked to complete a separate survey for each level, allowing for the possibility that raters may favor certain criteria over others depending on the level that they rate.

Once raters completed rating all of their assigned portfolios, they returned them to the writing coordinator who asked them to complete and submit the survey at that moment, so that their own rating process was still fresh in their minds. The survey results were entered into an Excel spreadsheet and prepared for analysis. The raters' ranking of

the criteria were entered into a matrix and then plotted in a bar chart graph along with the average ranking per level and overall for each criterion. The graphs were investigated to see if there were any discernable patterns among and between levels.

Each rater completed one survey for each level rated. There were 11 raters; some of them rated more than one level, resulting in a total of 22 completed surveys. Raters gauged the degree to which they felt the criteria influenced their rating. Rankings ranged between 1 and 5, with 1 meaning *not at all important* and 5 meaning *very important*. Averages were calculated for each level, as well as an all-level average.

Think Aloud Protocols

Six of the 11 raters were asked to participate in Think Aloud Protocols. These raters were selected by the ELC writing coordinator. Her decisions were based on her desire to have at least one TAP rater from each level and on her impressions of which raters would feel comfortable providing TAP data (and not see it as an attack or inquiry into their competency as a rater). Rater 19 was ill during the rating process, and although she was able to rate the portfolios assigned to her, Rater 20 replaced her as a Level 5 rater for the TAP investigation.

The TAPs took place during the second day of rating as soon as raters pick up their second batch of portfolios. When the selected raters arrived at the ELC to claim their second batch of portfolios, they were asked to provide TAP data before completing their second ratings. The decision to conduct the TAPs on the second day of rating was based on the expectation that on the first day raters would still be establishing their own rating process. However, by the second day, it was expected that raters would have created their own personal rating process, so a TAP session on the second day would be

more likely to provide an accurate sense of the thought processes that raters undergo when rating a portfolio.

Upon receiving their new portfolio groups, the selected raters individually met with the researcher for a TAP session. The researcher asked the rater to randomly select one portfolio from the rater's new portfolio group, and then rate the portfolio using the rater's normal rating process. The researcher asked the raters to voice aloud their thought process as they rated the portfolio, relating whatever internal questions or decisions naturally formed as they evaluated the portfolio. The researcher remained silent during the process except for neutral feedback such as "hmm, oh, uh-huh," etc. This backchanelling is suggested by Gass and Mackey (2000) who indicate that silence on the part of the researcher may make the subject nervous. Gass and Mackey warn that the researcher feedback should remain neutral; backchanelling should not be used to voice approval or dismay at the subject's remarks. The researcher made no other remarks, unless the raters became silent for long pauses, at which time the researcher reminded the raters to continue voicing their thoughts with comments such as, "Please continue" or "What are you thinking?"

Each TAP session (one per selected rater for a total of six sessions) was audio recorded and later transcribed. The transcriptions were reviewed in order to locate similarities or differences among raters. Attention was paid to several aspects of the rating process including how raters agree with, disagree with, or prioritize rating criteria. Another point of interest was how raters read through the writing samples as they decided upon a score. For example, whether raters graded each sample separately and then

averaged them for a final rating, or whether raters skimmed all samples and then decided upon a holistic score, etc.

Rater Interviews

The rater interviews took place a few weeks after the LAT rating. Four raters were asked to participate in these post-rating interviews. The purpose of the interviews was to clarify and further investigate the processes that raters use when grading portfolios. The interviews were conducted over e-mail in order to allow the interviewees time to thoughtfully consider and respond to the questions. The interviews were analyzed as were the TAPs: the responses were reviewed in order to locate patterns of common or discrepant behavior and attitudes among the raters.

Conclusion

This chapter has outlined the methodology used to gather data from several qualitative and quantitative sources: a Many-Facet Rasch Model (MFRM) analysis of exam scores, a rater survey, rater Think Aloud Protocols (TAPs), and post-rating interviews. The results from both the quantitative and qualitative analyses will help gather validity-related evidence from numerous sources. The purpose of collecting data from these sources will help make an argument concerning the scoring-related validity of the writing Level Achievement Test (LAT) at the English Language Center (ELC) of Brigham Young University (BYU). The results of these data analyses will be discussed in Chapter Four in preparation for a discussion of implications and conclusions in Chapter Five.

CHAPTER FOUR

Results

This chapter presents the findings of the Fall 2005 writing LAT validity study. First the results of the quantitative analysis are shown, followed by the results of the qualitative inquiries.

Quantitative Analysis

The quantitative analysis uses FACETS software, a Many-Facet Rasch Model (MFRM) tool that is based on item response theory (IRT). Item response theory differs from classical test theory in that it accounts for the interaction of exam facets on one another. FACETS software, for instance, reviews the test data and makes an initial estimate of the ability (for levels, classes, and examinees), severity (for raters), and difficulty (of writing criteria). Then, these initial facet estimates are reviewed for any unusual variances in each item. If an item varies in an inconsistent manner, it is assigned a higher infit mean square (MS); items with extreme variation are misfit and the software cannot compensate for their variation. These are problematic items that a researcher should review in order to improve test validity.

FACETS plots levels, classes, examinees, raters, and rating criteria on a single logit scale indicating the estimated ability (for level, class, and examinee), severity (for raters), and difficulty (for criteria) of the respective variables. It was expected that the five ELC levels should be equally dispersed along the proficiency scale with Level 5 at the top and Level 1 at the bottom; the same was expected of classes. Likewise, it was expected that higher level students be placed higher on the ability scale than lower level students; this would serve as evidence towards the LAT's validity. It was expected that

raters be tightly grouped around the mean, indicating high inter-rater agreement and serving as additional evidence of validity. Finally, it was unknown what location the rating criteria will take along the difficulty scale; a tightly clustered grouping would indicate that raters rely on all criteria equally. Conversely, a wide dispersion would indicate that raters rely on some criteria more than others when assigning a score.

IRT model estimation attempts to match all variable items to an expected model. Items that do not fit this expected distribution model are flagged at misfit. Although FACETS will indicate infit and outfit statistics, the difference between these two calculations only carries meaning when dichotomous data is analyzed. Because the LAT scores are polytomnous (a score along a 13-point scale), there is, therefore, little difference between the meaning of infit and outfit statistics; in this analysis they will be treated as synonymous. Items that are flagged as highly misfit (where the infit value for the item is greater than, or less than, the mean infit Mean Square plus or minus twice the standard deviation) are considered problematic (see Kim 2006). Too many problematic items casts doubt on the validity of the LAT.

The final Rasch-related evidence will come from IRT category response curves. FACETS will plot the proficiency distribution curves for ELC levels. A graph with evenly distributed levels will provide evidence towards validity; a graph with disordered or uneven level curves will cast doubt regarding the LAT's validity.

One of the advantages of MRFM analysis is that it graphs all facets on a single logit scale. Figure 4.1 provides a high-level overview of the data analysis. This table shows the estimated ability, severity, or difficulty of all items in each facet in relation to one another. In other words, the first column, *Levels*, estimates the ability of any given

student in each of the five ELC levels; a student in Level 5 (which is highest on the *Level* ability scale) is expected to have greater ability than a student in Level 4 (which is lower on the scale). The second and third columns, *Classes* and *Examinees*, also measure the estimated ability of students only this time at the class and individual level; classes and individuals appearing closer to the top of the column are estimated to have a higher writing ability than those near the bottom of the column. The forth column, *Raters*, represents the relative severity (at the top) or leniency (at the bottom) of the exam raters. The fifth column, *Criteria*, represents the guiding criteria that raters used to help them determine a portfolio's score. Criteria that raters graded more severely are found near the top of the scale (criterion difficulty) and items that raters were more lenient on are found closer to the bottom on the scale (criterion ease). In other words, criteria for which it was more difficult for students to receive a high score are near the top of the logit scale; easier criteria (i.e., those with higher student scores) are found near the bottom of the logit scale. The final facet column, *Scale*, represents the categories that raters used when grading portfolios; ability represented by a higher number (i.e. 7) is higher on the scale than a number representing a lower writing ability (i.e. 1). In a general sense, the information in Figure 4.1 allows a researcher to view the relative standings of analysis variables; any disordering of items within a variable could be seen. For example, if L4 appeared higher on the *Level* ability scale than L5, it would suggest that an average Level 4 student would have an estimated ability higher than an average Level 5 student. Disordering such as this could cast doubt on the validity of the exam.

In addition to the information in Figure 4.1, FACETS can generate additional measurement reports with more detailed information. This section describes the

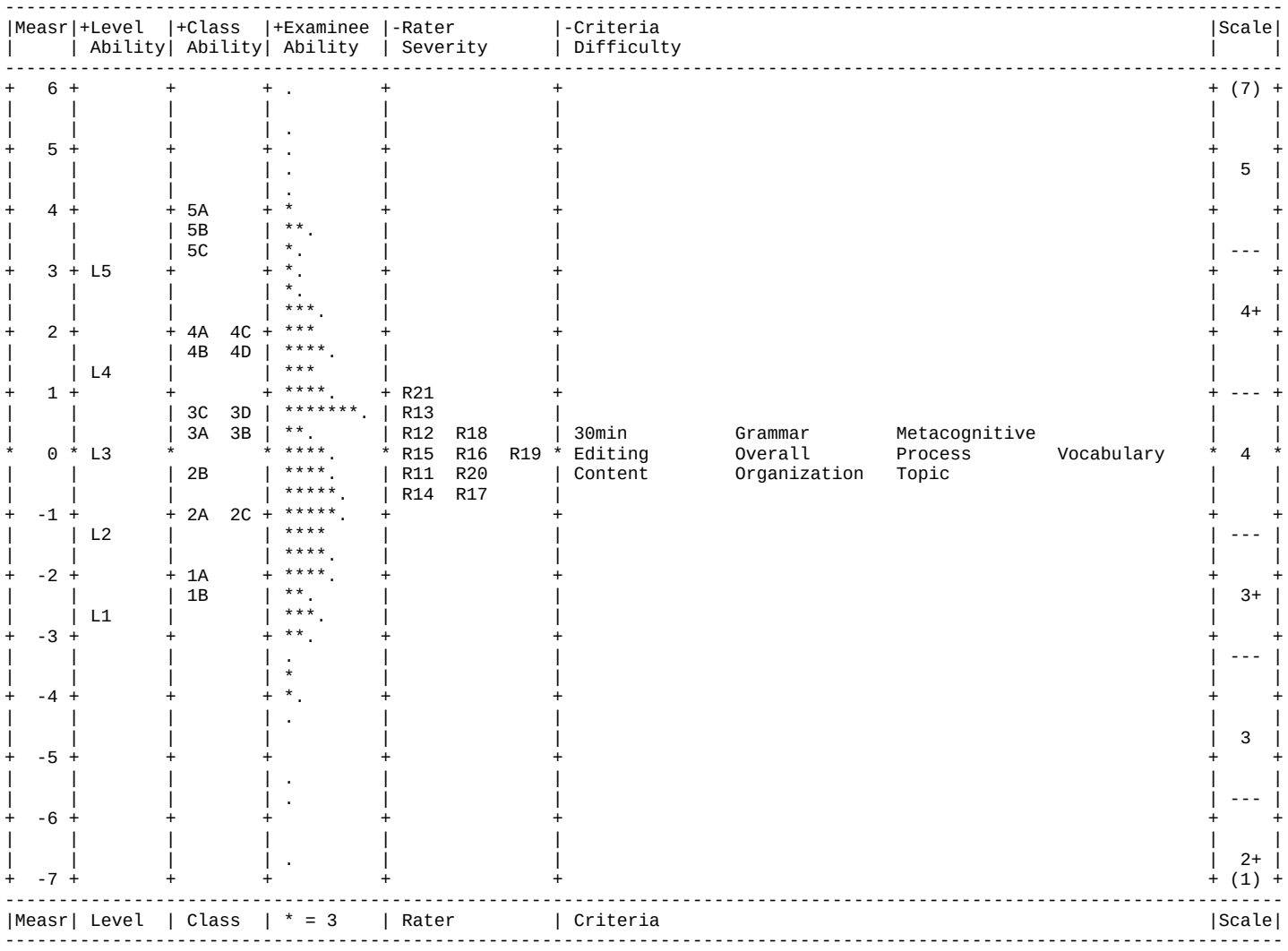


Figure 4.1 Summary of All Facets on Logit Chart

measurement reports for level, class, examinee, rater, writing criteria, and rating scale. It is followed by a report of the unexpected responses among the facets. Each of these reports is described as it relates to evaluating the effectiveness of the LAT rating process. Three measures of effectiveness are considered: ability ordering, fit statistics, and reliability separation index. The ability ordering is similar to the information in Figure 4.1 only at a more detailed view. All items should be properly ordered according to their logit values; items that are expected to be higher than others (according to ability, severity, or difficulty) should have higher logit values. Fit statistics is a measure of the degree to which items match the Many-Facet Rasch Model. Once FACETS has assigned logit values to each item, the software then reviews each interaction of that item to verify whether any items are acting in an unpredictable manner (e.g., a class in which some students perform exceedingly well while others score surprisingly poorly, or a rater who is inconsistently severe or lenient). Finally, the reliability separation index is a measure that describes the degree that items in a single facet are differentiating one from another. A high reliability separation index (1.0 is the highest value) is usually desirable in most facets; it shows that the exam is effective at separating items of that facet, for example, examinees or levels. However, in the case of raters, a low reliability separation index is usually preferred; this shows that raters function as a cohesive group. A high separation index for raters indicates low inter-rater reliability. However, the reliability separation index is not the same as more common inter-rater reliability measures. It tends to give a much more severe measure of reliability. As such, reliability index separation for raters should only be compared with other IRT reliability index separation indices and not traditional measures of inter-rater reliability.

Level, Class, and Examinee

Level and Class measurement reports (Appendices D and E respectively) indicate that levels and classes are ordered as expected. Higher levels and classes are estimated with a higher ability than lower levels and classes. There is no disordering among levels; however, some classes are estimated at a slightly higher ability than other classes at the same level. The gap between the highest and lowest classes at each level is much smaller than the gap between two adjacent levels. In other words, there is a greater difference in ability between the average Level 4 class and the average Level 3 class (1.75 – 0.47 = 1.28 logits) than there is between the highest and lowest Level 4 classes (1.91 – 1.59 = 0.32 logits).

The standard deviations for levels and classes are very small, so this indicates that individual levels and classes tend to function as a group. Also, there are no misfit values which indicates that each level and class is performing in a consistent manner. Both levels and classes have a reliability of separation index of 1.0 which indicates that the LAT is doing a good job of separating among levels; the rating process appears to be effective at differentiating between levels.

The MFRM analysis reports that, in general, examinees are ordered as expected. Examinees in lower level writing classes have a lower estimated ability value; those in higher level classes tend to have higher estimated ability values. However, some misordering is present; some students of lower levels performed better than students of a higher level. Those level lower examinees who scored higher than examinees of a higher level (and vice versa) are indicative of students whose estimated ability is higher or lower

than expected. The MFRM analysis will compensate for rater severity/leniency in these cases, so the estimated examinee ability is rater-independent.

Table 4.1 shows a sample of examinees along with their logit value, standard error, and infit mean square. The six-digit Examinee ID in the first column is coded with the following system: the first two digits indicate level (50 = Level 5, 40 = Level 4, etc.), the next two digits represent class (a random number assigned to each class in a level), and the remaining two digits indicate an individual student (a random number given to each student in a class). Using this knowledge of the Examinee ID, it can be seen that the second column demonstrates how overall there is some misordering of examinees. Examinee 401220 in Level 4 received an ability measure of +1.89 which is higher than the ability assigned to Level 5 Examinee 501412 (+1.86). However, this misordering is very small and could be the result of error. Table 4.2 shows additional selected examples of students who are misordered.

Table 4.1

MFRM Measurement Report for Misfit Examinees

Examinee ID	Ability (logits)	Standard Error	Infit MS
501301	+2.30	0.34	2.04
501401	+1.99	0.35	2.36
501412	+1.86*	0.36	2.01
401220	+1.89*	0.28	3.26
401017	+1.02	0.34	2.37
401217	+0.93	0.28	1.85
300717	+0.72	0.34	2.18
200504	- 2.75	0.33	3.16

*indicates misordering in examinee ability among proficiency levels

Fit statistics for examinees show that some examinees are misfitting. Lynch and McNamara (1998) and Park (2004) provide a means for interpreting misfit. They suggest that items are misfit if their infit value is greater than or less than the infit MS mean plus

twice the standard deviation. In the case of the examinees in this study, a student will be considered misfit if that student's infit value is greater than 1.82 or less than -0.24 (0.84 +/- 2(0.49)). Eight examinees fit this definition, and all are listed in Table 4.1. They account for 3.18% of all examinees, an acceptably low value as discussed by Kim (2006). These examinees are performing unexpectedly as compared to their peers of the same level, class, or rater. They may have an extreme ability higher or lower than the MFRM model can account for, or their performance on the different writing criteria is highly unusual. In short, their performance on the exam varies from the expected model.

Table 4.2

MFRM Measurement Report for Selected Misordered Examinees

Examinee ID	Ability (logits)	Standard Error	Infit MS
300606	+2.12	0.28	0.78
200403	+0.34	0.33	0.44
501504	+0.65	0.36	0.67
100112	-0.63	0.33	0.62
401105	-0.98	0.34	1.61
300616	-2.81	0.33	0.30

It should be emphasized that there is a difference between misodered and misfit. When it comes to examinees, some misordering may occur: students in lower levels may perform higher on the logit scale than students in higher levels. This occasionally can happen when students of exceptional writing ability are placed in a low level and when students of struggling ability are placed in a high level. This is especially possible at IEPs (intensive English language programs) where students are placed in skill classes all of the same level. Although some students may be exceptionally weak or strong in writing as compared to their other language skills, they are placed in a level based on their average ability. This inevitably results in misordering among examinees since some higher level

students may have weaker writing skills than their lower level counterparts. Misfitting, however, is a far greater problem. Misfitting examinees are those who perform in unpredictable ways and may indicate cheating, misunderstanding, or guessing. Misfitting examinees may also result from other intervening misfit variables such as examinees graded by misfit raters or with misfit criteria.

The standard error for examinees is high (0.34), mostly likely due to only double-rating the portfolios (if more rating were used, or there was higher inter-rater agreement, then the standard error scores would be lower). The reliability separation index is 0.98 which indicates that the exam does an effective job of differentiating students into varying ability measures. A complete MRFM measurement report table for examinees can be found in Appendix E.

Raters

The measurement report generates a variety of measures that can reveal useful information about the performance of the LAT raters (Table 4.3). First is the degree of rater severity. The second column, *Severity*, indicates that the span between the most severe rater, R21 (+0.89), and the most lenient rater, R14 (-0.57), is 1.46 logits. Although this span is greater than the 0.51 severity span reported by Sudweeks, Reeve, and Bradshaw (2005), it is relatively smaller than the 3.31, 2.43, and 5.24 spans reported by Bachman, Lynch, and Mason (1995), Lynch and McNamara (1998), and Park (2004) respectively. A large rater severity span indicates that the reliability of an assigned score is more likely to vary depending on which rater grades the student work. If the span is small, then there is a greater chance that a student's score remain the same regardless of

which rater grades the portfolio. FACETS analysis will account for discrepancies in rater severity so long as raters are consistent (i.e., not misfit).

Additionally, MFRM also provides a reliability estimate with the rater measurement report. This reliability of separation index is not an inter-rater reliability measurement, but instead represents the degree to which raters act independently of one another. The closer that the value is to 1.0, the more likely raters differ in their degree of severity/leniency. A value of 0.0 would indicate that raters have no separation in their degree of severity/leniency. In other words, the reliability of separation index for raters indicates the degree of unwanted variability among raters. The reliability of separation index for this study appears relatively high at 0.97. However, this is comparable to the 0.84 index reported by Sudweeks, Reeve, and Bradshaw (2005), the 0.97 index reported by Parks (2005), the 0.92 index reported by Bachman, Lynch, and Mason (1995), and the 1.00 index reported by Lynch and McNamara (1998).

Table 4.3

MFRM Measurement Report for Raters

Rater ID	Severity (logits)	Standard Error	Infit MS
R20	- 0.37	0.07	0.53
R16	- 0.09	0.07	0.55
R12	+0.32	0.06	0.57
R18	+0.18	0.10	0.63
R21	+0.89	0.10	0.72
R17	- 0.51	0.07	0.72
R13	+0.55	0.07	0.92
R14	- 0.57	0.07	0.92
R15	+0.12	0.08	0.95
R11	- 0.38	0.06	0.98
R19	- 0.12	0.07	1.90
Mean	0.00	0.07	0.85
Standard Deviation	0.46	0.01	0.39

Reliability of separation index = 0.97

A more important issue is rater consistency. The *Infit MS* (mean square) column indicates the degree to which raters followed a consistent pattern in their use of severe or lenient grading. As stated earlier, Lynch and McNamara (1998) and Park (2004) provide a means for determining misfit. In the case of raters, they state that raters are misfit if their infit value is greater or less than the infit MS mean plus or minus twice the standard deviation. In the case of this study, a rater will be considered misfit if that rater's infit value is greater than 1.63 or less than 0.07 (0.85 +/- 2(0.39)). Only one rater fits this definition, R19 with an infit score of 1.90.

Writing Criteria

There appears to be little variation in the rating of writing criteria as seen in Table 4.4. The criterion with the highest difficulty value is the 30-minute essay sample (+0.49 logits); the criterion with the lowest difficulty value is topic (-0.49 logits). The difficulty span is less than one logit, so all criteria are closely centered around the overall rating score. This analysis does not account for any differences in criteria difficulty across levels. Instead, it indicates that the 30-minute essay, grammar, the metacognitive essay, editing, and vocabulary are more difficult criteria; topic, organization, content, and writing process are easier criteria. In general, raters do not assign individual criteria scores that deviate far from the overall rating.

Fit statistics for writing criteria do not reveal any concerns. Criteria are misfit if the Infit MS is greater than 1.14 or less than 0.58 (0.86+/-2(0.14)). Only one criterion fits this description: overall score. This indicates that the overall score varies in an unexpected manner. This could be due to the holistic scoring of the exam; the overall score is not necessarily an average of the analytic criteria, so it may not fit the model as

accurately when compared to the other criteria. The reliability of separation index for this facet is high at 0.96. So, although the difference in difficulty on the logit scale indicates that the criteria may not differ greatly, the reliability of separation index suggests that raters appear to be able to discriminate among the criteria more effectively than first indicated.

Table 4.4

MFRM Measurement Report for Writing Criteria

Criteria	Difficulty (logits)	Standard Error	Infit MS
30-minute essay	+0.49	0.07	1.05
Grammar	+0.39	0.07	0.92
Metacognitive essay	+0.29	0.07	1.00
Editing	+0.15	0.07	0.85
Vocabulary	- 0.02	0.07	0.83
Overall	- 0.05	0.07	0.52
Process	- 0.15	0.07	0.85
Content	- 0.17	0.07	0.86
Organization	- 0.43	0.07	0.78
Topic	- 0.49	0.07	0.90
Mean	0.00	0.07	0.86
Standard Deviation	0.33	0.00	0.14

Reliability of separation index = 0.96

The most valuable data from the MRFM analysis of writing criteria is separation of criteria into difficulty categories. As indicated earlier, Table 4.4 lists the criteria in order of difficulty from most to least difficult. The most difficult criteria are the 30-minute essay, grammar, the metacognitive essay, editing, and vocabulary. The least difficult criteria are writing process, content, organization, and content. The easiest criteria, then, are factors that are commonly grouped as global issues in writing, criteria that could be described as macro-linguistic features. If raters are grading both the 30-minute and Metacognitive essays as samples of writing fluency and accuracy, then the most difficult criteria could be grouped as local issues in writing, micro-linguistic

features. This separation of criteria into these two groups of greater/lesser difficulty may indicate areas of student and teacher strength/weakness.

Rating Scale

The MRFM analysis report for the rating scale indicates that, in general, the scale is appropriately used by raters. When evaluating a rating scale using MFRM, there are a few key measurements to analyze. First, a researcher should verify whether the step calibration values are properly ordered. Step Calibration values refer to the midpoints between two categories. For instance, the +12.55 step calibration value refers to the graphical location along the logit scale where the probability curves for 7 and 6+ intersect. At the midpoint, the likelihood of a student at that ability level scoring 7 or 6+ is equal. Because midpoints refer to the midpoint of two category curves, there is one less step calibration value than there are scale categories. The second column of the rating scale measurement report in Table 4.5 reveals that this ideal has been achieved: there is no disordering of scale category step calibration values.

Table 4.5

MFRM Measurement Report for Rating Scale

Category	Step Calibrations	Counts	Percentage
7	+12.55	37	0.7%
6+	+10.85	150	2.9%
6	+ 8.51	344	6.7%
5+	+ 5.80	501	9.7%
5	+ 3.46	766	14.9%
4+	+ 1.16	809	15.7%
4	- 1.36	824	16.0%
3+	- 3.35	550	10.7%
3	- 5.74	506	9.8%
2+	- 7.71	348	6.8%
2	- 11.85	285	5.5%
1+	- 12.32	25	0.5%
1	--	5	0.1%

In addition to verifying correct ordering of step calibration values, a researcher should also analyze the scale for distribution measures. This can be measured by comparing the count or percentage values for the scale categories. Ideally, there would be an equal distribution in each category. In reality a more normal distribution is expected: higher distribution at the middle categories (2+ to 5+) and lower distributions towards the tail categories (1 to 2 and 6 to 7). In general, this is the case; there is a fairly uniform distribution of ratings in the middle range (3 to 5+) and tapering at the end points. The data from the third column of Table 4.5 is presented in visual form in Figure 4.2.

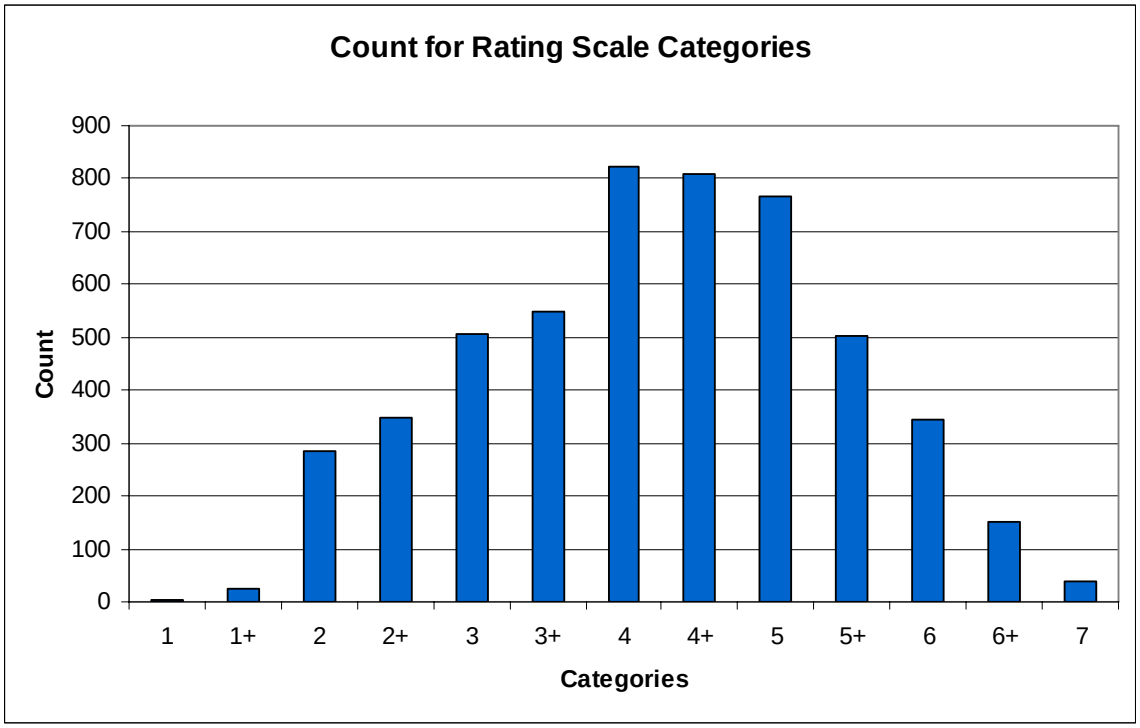


Figure 4.2 Graph of Count for Rating Scale Categories

Another way to investigate the category distribution is to view the probability curve graph. Ideally, the graph should show uniformly rounded “hills” for each category, equally spaced along the axis. Figure 4.3 shows the probability curve graph for this study.

In general, there is uniform distribution of curves. However, the curves for 1+ and 6+ are relatively low. Also, the curve for 2 is especially high. This indicates that raters may have trouble distinguishing between categories in the 1 to 3 range; they may also have difficulty with 6 to 7 range.

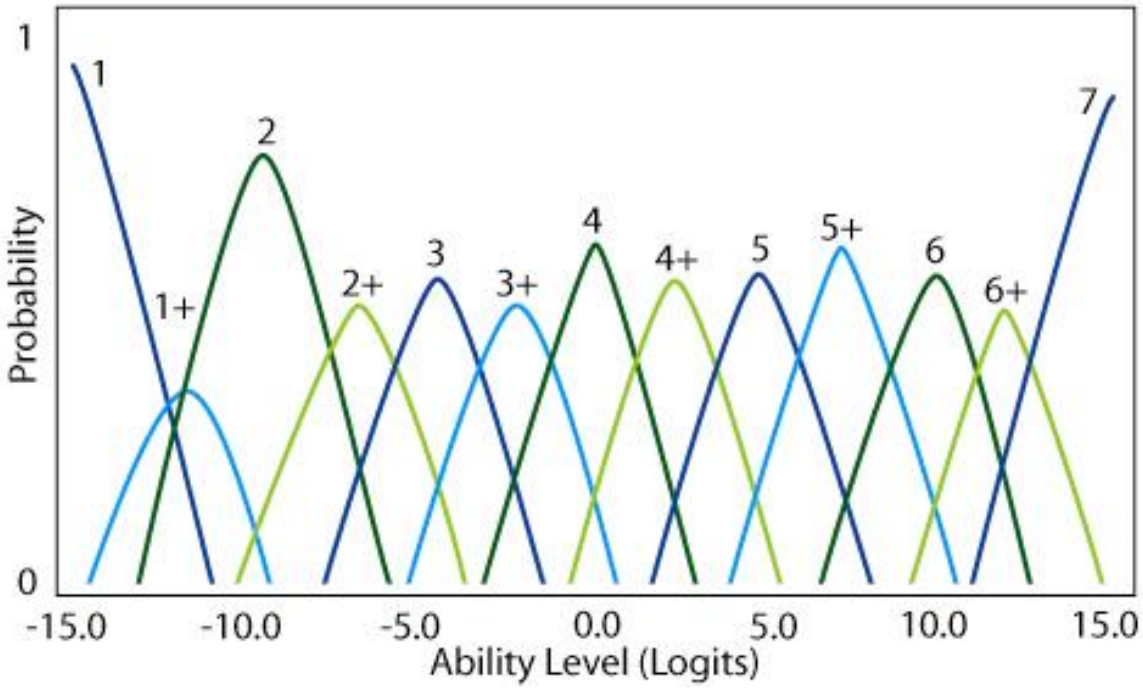


Figure 4.3 Probability Curves for Rating Scale Categories

Qualitative Analysis

The qualitative results for this study provide additional support for the quantitative findings and place them in context. Qualitative results come from three sources: a rater survey, rater Think Aloud Protocols (TAPs), and rater interviews.

Rater Surveys

The purpose of the rater survey was to gather information regarding which rating criteria the raters perceived themselves as favoring when assigning a portfolio score. Due

to the low number of responses, this survey is not meant to be a representation of any given rater's preferences at any given level. Instead, the results of this survey are designed to provide a general indication of rater preferences. The results of the survey are simply meant to suggest possible trends among rater self-perception and preference. The data indicates that there are some differences between overall rater weighting of criteria, as well as differences among levels.

Figure 4.4 represents the average score for each criterion per level, as well as the all-level average. These results indicate that some criteria were valued by all raters at all levels (length of essays, essay organization, accuracy of grammar use, and the 30-minute essay sample). Other criteria were more important when rating lower levels (correct spelling, correct punctuation, and the metacognitive essay sample) or when rating higher levels (difficulty of topic, depth of topic, vocabulary use). A number of raters indicated that they felt that additional concepts should be added to the rating criteria. These included cohesiveness of ideas, ability to analyze, voice, and sophistication.

Think Aloud Protocols

Six raters participated in Think Aloud Protocols (TAPs), one from Level 1, one from Level 2, two from Level 3, one from Level 4, and one from Level 5. The TAPs were all conducted during the scoring of the second batch of rating. Raters had already scored one or more groups of portfolios and had established an internal process for rating. The purpose of the TAPs was to gather information regarding this internal thought process that raters undergo when assigning a score to a portfolio. Raters were asked to voice aloud their thoughts as they rated one portfolio. The monologue was digitally recorded

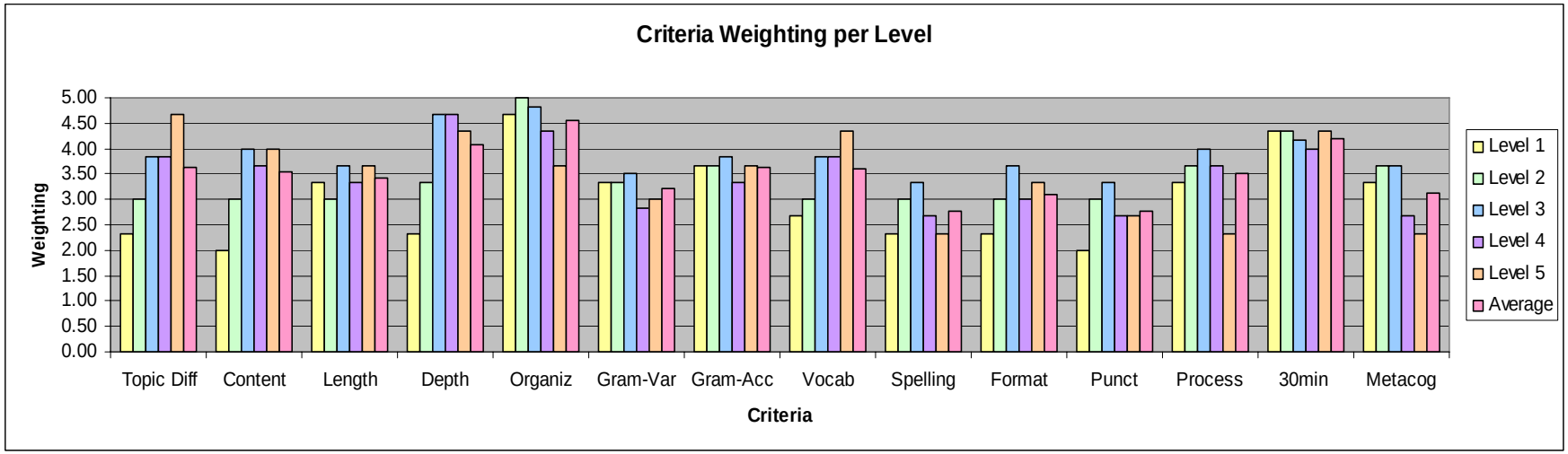


Figure 4.4 Graphical Averages of Criteria Priorities Based on Rater Survey Results

with the researcher present. The recordings were then transcribed one week later. Full transcription of the TAPs can be found in Appendix F.

The transcripts were analyzed for any differences or similarities among the processes employed by raters. In general, all raters followed a process similar to the suggested process demonstrated at the rating calibration/training meeting: raters read through the writing samples, making preliminary judgments about the holistic score; they adjust their decision based on any remarkable use of the writing criteria; then they assign a holistic score and provide analytical feedback that more or less averages the holistic score. This general process was common among all TAP raters. Any given rater would read a writing sample, comment on the rating criteria, assign an interim score based on one same, and then continue on to the next sample, repeating the process and readjusting the interim score if needed. In general, raters followed this pattern for each sample until all samples had been read. Then a holistic score was assigned.

However, there were also differences among the processes rater used, including the order in which writing samples were read, the criteria that raters favored, and the time raters spent on a portfolio. Although most raters read the writing samples in the standardized portfolio order (Essay 1, Essay 2, Metacognitive Essay, 30-minute Essay), Rater 17 read the 30-minute essay first. She explained her decision saying, “I like to start with the 30 minute because then you get an idea what their writing is like on their own.” So although most raters follow the standard order to dictate their reading order, Rater 17’s process indicates that some raters may follow a different order when reading portfolio samples.

Raters tended to pay attention to many of the same criteria when rating (grammar usage, sentence construction, level-appropriateness of topic, organization). They also appeared to agree on how to judge the students' understanding of the writing process. Raters were lenient on students who made few changes throughout their drafts if those students did not receive sufficient feedback from teachers or tutors. For example, Rater 15 noted the lack of organizational and content changes in a Level 3 student's drafts. He said, "In the drafts, looks like she [the teacher] made some suggestions. Mostly it was grammar or microlinguistic feedback. She [the student] didn't make many big changes, but it looks like she didn't get any feedback on that." Rater 17 was lenient on a Level 4 student based on the same reason: lack of organizational feedback on the drafts. She said, "Looking at his drafts… there's a lot of grammar and vocabulary help. His first draft is two paragraphs long, which means he didn't finish it. And he didn't get much feedback on it." She explicitly explained her justification for being lenient, saying, "I'm going to look at the teacher comments on the drafts to see if he revised, or if these problems were never addressed then it's less their fault if no one ever helped them."

These comments suggest that the grading of the writing process may be as much a measure of the teacher or tutor as it is of the examinee. If a teacher or tutor provides no feedback, then a rater will not alter a student's score; however, if a draft contains teacher or tutor feedback and a student does not follow that advice, it is possible that a rater will penalize the examinee.The issue of teacher/tutor feedback weakens the validity of the LAT. If raters are basing their assessment on revisions as instructed by teachers or tutors, then it is possible that the LAT is partially a measurement of teacher or tutor ability and not just examinee ability.

At the same time, raters also wanted to verify that students did follow advice that their teachers provided. Rater 14 repeatedly looked for this when reviewing a Level 3 paper. She remarked, “Let’s check to see if they followed teacher’s comments in improving their drafts… yes… good… this one [draft] looks good.” Later, she made a similar comment, “Now let’s check to see if they followed the teacher’s advice for revision… yes… good.” When judging students’ understanding and use of the writing process, raters appeared to make allowances when teacher feedback was missing, but expected students to make changes when that feedback was present.

Although raters equally valued some criteria, they differed in their focus on other aspects of writing. For example, some raters placed a greater emphasis on the metacognitive and 30-minute essay samples than other raters. As mentioned earlier, Rater 17 used the 30-minute essay as an introduction to the student’s writing ability; other raters used that sample as a confirmation of the interim score. For example, while reviewing the 30-minute essay at the end of the rating process, Rater 12 said, “I think the 30 minute essay reflects the skills he showed in other papers and in Thinking About My [metacogitive] Essay.” Here the 30-minute essay was used to help the rater feel confident that she had assigned the correct score. Rater 14, however, used the 30-minute essay as an adjustment to the interim score. She had tentatively been thinking of assigning a high score to the portfolio after reading the first two sample essays. This changed once she read the metacognitive and 30-minute essay. She said, “The metacognitive was not as clear. Deserves a 4…” Then after reading the 30-minute essay, she came to the same conclusion, “I think that it deserves a 4.” Both of these scores differed from the final holistic rating that she gave the portfolio. She determined the final score, saying, “Okay,

this was definitely a very strong portfolio. Not quite an Honors [score of 5], but definitely a high pass [score of 4+] because of its diversity, clarity, and organization." Clearly she was impressed with the first two writing samples, but because the last two writing samples were of poorer quality, she decided to assign a slightly lower score to the portfolio.

Formatting is another criterion that was not universally important to all raters. Rater 12 is the only rater who made mention of formatting. She noticed problems with a Level 1 essay saying, "I wish this paper were typed, but since this is Level 1, I won't be paying attention to that." Although it appears that this did not influence her score, she later remarked on the formatting again. She noted, "But the paper is not well formatted, because at the beginning it should have been indented, every paragraph." It is possible that formatting issues may have played a role for other raters as well, but there is no evidence of that in the TAPs.

Lastly, the TAPs revealed that raters differed in the amount of time they take to grade a portfolio. Table 4.6 shows the recorded times that the six raters took to grade their assigned portfolio. There is no discernible pattern in the data; it simply indicates that raters vary in the amount of time they take to review a portfolio. There is currently no time restriction, or even guideline, for raters.

Rater Interviews

Rater interviews were conducted several weeks after the rating process. This was done as to allow the researcher time to analyze the data and ask any follow-up questions that might help clarify trends/anomalies in the data or provide further insight into the

Table 4.6

Time Taken by Raters to Review a Portfolio

Rater ID	Level	Time (minutes:seconds)
R12	1	07:23
R13	2	14:40
R14	3	09:32
R15	3	07:03
R17	4	10:39
R20	5	02:20
Average	--	08:36

rating process. Four raters were asked to respond to four open-ended questions about the effectiveness of the rating process:

1. How does your LAT rating process differ from your essay grading process?
2. Do you feel that the LATs are a fair assessment of student ability? Why or why not?
3. How effective do you feel the LAT calibration meetings are? Do you have any suggestions for the improvement of these meeting?
4. Do you have any other suggestions for the improvement of the LATs?

In the semester following the data collection for this research project, the ELC writing coordinator made some changes to the writing program in order to improve the quality of the LATs. These changes, such as improved calibration and teacher coordination, were based on rater feedback and an initial analysis of the data for this research project. As such, when the rater interviews were conducted, some changes to the grading process had already been underway. The raters' responses to the interview questions reflect their suggestions for the new semester, their opinions about recently implemented changes, and their expectations about how these changes will affect the writing LAT in the future.

When asked if their LAT rating process differed from the essay grading that they do throughout the semester, raters indicated that there are some key differences. First and foremost, some raters acknowledge the higher stakes, and yet more rushed process, of LAT rating. Rater 13 said, “My LAT rating process was much more stressful because of the time pressure I was under. Because I felt like I had to get it done quickly, and because I didn't know some of the students I would have felt bad rushing through it and leaning toward the side of a lower score.” He expressed his awareness of the responsibility that a rater’s role has on student grading. Although Rater 17 likewise recognized the difference in grading between LAT rating and classroom grading, she did not feel the same anxiety as Rater 13. She explained, “I read a lot faster through the LATs because there are more essays and I'm just looking for an overall, holistic score. Also, I don't mark [provide feedback on] the LATs at all, so that goes faster. Oh, and because I may not know the student and it's a summative evaluation (rather than the more formative ones during the semester), I tend to grade them more formally.” For some raters, the shorter time frame for LAT rating, in comparison to formative essay grading, can cause stress. Others feel comfortable in this role, citing the summative role of LAT rating as a reason to approach the rating process in a faster, more holistic manner.

Raters 16 and 11 claimed that their LAT rating process is very close to their semester-long grading process. Rater 16 explained, “In order for me to rate well I follow the same process I use in rating essays in class so that my rating is fair and consistent. I follow the same thinking pattern and it makes it easier for me too.”

Rater 11 made similar remarks when he detailed how he made changes to his classroom teaching following the LAT process in question. He noted, “This semester I've

implemented the LAT rating process in the way I grade my students' essays… In particular, I've incorporated the areas that the LAT focuses on and adjusted them slightly to match more with what I'm teaching and expecting my students be able to do. I also added values for each section in order to emphasize to the students which areas I believe, and hopefully, what I hope they soon will understand, are vital areas in improving their overall writing ability. I've used the LAT rating criteria to evaluate my students' in-class and 30-minute timed essays, so that they will become familiarized with how they'll be graded on the final writing LAT portfolio. I even have them apply a TOEFL-based criteria for self- and peer-reviewing their essays." The approach described by Rater 11 indicates that some, if not all, raters are conscious of the LAT rating throughout the semester and attempt to teach and grade their students so that the LAT process will be more natural for both student as well as teacher-rater.

When asked whether they felt the LATs were a fair assessment of student writing ability, raters gave mixed results. Raters 14, 16, and 11 indicated the need for a mandatory writing sample for each level. They felt that students who included more challenging writing samples in their portfolios (such as a research paper) were more harshly graded than students who chose to include samples that did not require them to exercise writing skills that stretched their ability. Rater 16 replied that "one of my lower students got a higher grade for the LATs than my more proficient students because he put in shorter, easier papers and was rated accordingly and my other student put in his academic paper which warranted more severe grading." This is another area that weakens the validity of the LAT. If students who include easier assignments in their portfolios are

graded more lenient, then it is possible that the LAT is not just a measurement of student ability, but is also based on the type of writing samples.

Raters' responses to this question also revealed a difference in opinion about the other writing samples. Rater 11 questioned the validity of the 30-minute writing sample saying, "I cannot clearly see the assessment value of the 30-minute timed essay… I know that in the level objectives that the students of each level are required to write so many words in 30 minutes, but this does not really support with our teaching students to apply the steps of the writing process in becoming better writers. I've had students in all levels apply the process and produce incredibly well-thought, well-organized, coherent essays in class, and then freeze up with test anxiety when taking a timed essay exam. For some students, their ability and skill in writing in-class and timed essays, may match up somewhat equally, but there are others where it does not. This is why while I assess the writing LATs, I look at the timed essay very last rather than the first." Rater 17, on the other hand, takes the opposing view. She doubts the validity of the multi-draft essay samples explaining, "LATs assess their ability to write a multiple-draft essay, and they can receive help from many sources. Their 30-minute and metacognitives sort of show their writing ability without help, but it's probably not quite enough." These comments serve as evidence that there is disagreement among raters regarding the priority of writing samples in the portfolio.

In all four interviews, raters expressed their appreciation of, and reliance on, the pre-LAT calibration meeting. They stressed the value of these meetings citing effective benchmarks, inter-rater discussion, and multi-level grading awareness as the most important aspects of the calibration sessions. Raters 17, 11, and 16 all expressed the

desire to conduct inter-rater calibration earlier in the semester so that teaching, rating, and assignment creation were all more aligned.

In the final interview question, raters were asked if they had any suggestions for improving the LAT. Once again, raters cited the need for a required paper for each level that challenged the most advanced writing skills of that level (i.e. an academic research paper for Levels 4 and 5). Rater 11 also suggested that raters be made aware of research studies such as this one so that they can self-assess their rating effectiveness and compare their processes to that of their peers. See Appendix H for complete responses to the rater interviews.

This chapter has summarized the results of both the quantitative and the qualitative analyses. Data was collected from various sources including exam scores as well as rater surveys, rater Think Aloud Protocols, and rater interviews. The implications of these results will be discussed in the following chapter.

CHAPTER FIVE

Discussion and Conclusion

The purpose of this chapter is to answer the research questions, and then to discuss implications, limitations, and suggestions related to this study. First, the research questions are answered using data from both the quantitative and qualitative analyses in order to provide a more complete understanding of the results. This is followed by sections dedicated to suggestions for improving the effectiveness of the LATs, teaching implications, and limitations of this study. Finally, suggestions for further research are shared before final conclusions are offered.

The research questions that guided this study are:

Quantitative evidence:

1. How valid are the writing LAT scores based on a Many-Facet Rasch Model analysis? This question is further separated into the following questions.
    a. How well do the LATs distinguish among levels, classes, and examinees?
    b. How severe are the raters in relation to one another?
    c. How consistent are the raters?
    d. How difficult are the writing criteria in relation to one another?
    e. How well is the rating scale used?

Qualitative evidence:

2. What is the degree of rater agreement on the priority of rating criteria among and between levels?

3. How do raters apply the rating criteria to determine a LAT score? (i.e. prioritizing some criteria more than others, and valuing some criteria more in higher levels than in lower levels, etc.)
4. How do raters use portfolio samples to negotiate a LAT score? (i.e. holistically based on all scores, or an average based on individually determined scores for each sample, etc.)

Discussion of Results

1. *How valid are the writing LAT scores based on a Many-Facet Rasch Model analysis? This question is further separated into the following questions.*

The qualitative evidence from the MFRM analysis suggests a high degree of scoring-related validity in the writing portfolio LAT. Specific portions of the analysis are discussed below.

a. How well do the LATs distinguish among levels, classes, and examinees?

Levels, classes, and examinees are performing as expected. Level 5 is higher on the logit scale than levels 4, 3, 2, and 1. Each level is evenly spread indicating that the exams differentiate levels into distinguishable groups. Classes of the same level are tightly grouped to one another with higher level classes performing better than lower level classes. For the most part, examinees are placed as expected along the logit scale. Although there is some misordering of examinees (i.e., Level 3 examinees performing better than Level 5 examinees), this is not unexpected given that some examinees from Level 3 may have superior writing ability to some higher level students. Even so, misordering is at a minimum. Overall, the MRFM analysis suggests that the writing

portfolio exam has a high degree of validity in placing and distinguishing among levels, classes, and examinees.

b. How severe are the raters in relation to one another?

In comparison to similar MFRM studies, this study suggests that the LAT raters have an acceptable level of severity/leniency span; there are no raters that are too lenient or too severe. However, if greater inter-rater agreement is desired, increased calibration and discussion could help.

c. How consistent are the raters?

Rater consistency is also within acceptable levels, with the exception of Rater 19 whose inconsistency could not be compensated for with FACETS software. Under normal circumstances, this would indicate that Rater 19 should either be removed as a rater or should received additional rating practice. However, as mentioned in the previous chapter, Rater 19 was ill during the rating process and this may account for her inconsistent rating behavior. The LAT administrator (i.e. the writing program coordinator) may wish to monitor Rater 19's performance in future semesters to see whether this inconsistency was an anomaly or is indicative of her usual rating process. Overall, the MRFM study suggests an adequate degree of rater validity; however, steps could be taken to improve the performance of raters.

d. How difficult are the writing criteria in relation to one another?

Despite the overall encouraging results to the above questions, the FACETS analysis raises some concerns, most notably the analysis of writing criteria. The scoring of the individual writing criteria is tightly clustered around the overall score. This suggests that the rating of individual writing features is not differentiating very well

among students. Instead, it appears that raters are selecting scores for individual criteria that are the same as, or are very close to, the overall score they assign a portfolio. As a result, individual criteria scores may carry little meaning and may not be a valuable source of feedback to students, nor do these scores provide any useful feedback to administrators and teachers about student ability in the individual criteria areas.

There are two reasonable explanations for this clustering of criteria scores. First, in rater training meetings, teacher-raters had been taught to use the rating criteria as guiding details as they evaluated the samples and assigned a holistic score. Then, raters were instructed to return to the feedback sheets and select scores for individual criteria that "more-or-less" averaged the overall score that had already been determined. As such, raters may feel pressured to select criteria scores that do not deviate remarkably from the overall score. In some cases this has resulted in raters who select criteria scores that are exactly the same as the overall score; the deviation is zero.

The second reason why there is little variation among criteria scores is a result of the feedback criteria sheets. The sheet limits raters to selecting criteria scores that are within two points of the expected level score. For example, a rater grading a Level 2 portfolio will use a feedback sheet with a scoring range of: 2, 2+, 3, 3+, and 4. This limits the degree of variation that a rater can assign to the individual criteria; scores are inevitably tightly clustered around the overall score even if the rater feels that there is a severe deviation in ability of one criteria over another. As a result, this analysis suggests that students may benefit from a more detailed or alternate form of criteria feedback.

Even though the criteria are tightly clustered, it should be reemphasized that criteria related to global/composition issues (content, organization, topic, and process)

were easier than local/language issues (grammar, editing, vocabulary, Metacognitive essay, and 30 minutes writing sample). This suggests that students understand and perform well on global issues but require more help at improving their performance on local/language issues. The writing coordinator may wish to improve the teaching of local/language issues in the classroom in order to improve student performance in these areas.

e. How well is the rating scale used?

Overall the MFRM analysis indicates that the rating scale is well used. Only one portion of scale was under/overused: the 1, 1+, 2 range. This weakens the rating scale's validity and suggests that this portion of the scale be reorganized to ensure a more uniform distribution of scores in the FACETS analysis. However, because this scale will be used in future semester when more lower level students may study at the ELC, it is wise to keep this section of the scale and reevaluate its effectiveness in future semesters.

2. What is the degree of rater agreement on the priority of rating criteria among and between levels?

The results of the rater survey, rater TAPs, and rater interviews indicate that there is variation among which criteria teachers of each level prioritize. Criteria received varying endorsements in the rater survey. For example, the 30-minute essay sample and organization were highly prioritized, but grammar, spelling, and punctuation received lower endorsements. Additionally, there is variation within each level. Some criteria are universally valued, such as organization, yet raters disagree on the importance of criteria such as vocabulary, topic, and self-reflection. This indicates that raters agree on the importance of some criteria, yet there is no consensus on the role that other criteria

should play in the rating process. This disagreement among criteria weakens the LAT's validity. Attention to be given to understanding why this disagreement exists and whether each level should have a different list of criteria based on differing writing needs of students across in different levels. Greater consensus among raters will increase the scoring-related validity of the LAT.

3. *How do raters apply the rating criteria to determine a LAT score? (i.e., prioritizing some criteria more than others, and valuing some criteria more in higher levels than in lower levels, etc.)*

As revealed in the response to research question 2, raters vary in their prioritizing of criteria. This inevitably leads to variation in the manner in which raters apply the criteria to grading a portfolio. Some raters rely more heavily on certain features than others. The difference of opinion in criteria prioritizing could account for some of the discrepancy between rater severity/leniency in the quantitative analysis. If raters are prioritizing different parts of the same essay (topic versus grammar), or different samples of the same portfolio (i.e., multi-draft essays over the 30-minute sample) then it is likely that they will assign a different score if they feel that there is an uneven performance of those criteria.

As Rater 17 pointed out in her interview, the 30-minute essay is the only real evidence that raters have of an examinee's fluency. Multi-draft essays are developed over time and, in some instances, are heavily influenced by a student's friends, family, tutor, or even classroom teacher. The rater survey results indicate that raters highly value the 30-minute essay sample, and yet the MFRM analysis shows that they grade it more severely than most other criteria. Although Rater 11 was the most skeptical of the 30-

minute essay, even he defended its inclusion in the LAT portfolio both as practice for the iBT (internet-based TOEFL) and as a sample for gauging independent, spontaneous writing ability. The 30-minute essay appears to be an important part of the LAT rating, though more could be done to investigate its usefulness.

4. *How do raters use portfolio samples to negotiate a LAT score? (i.e. holistically based on all scores, or an average of based on individually determined scores for each sample, etc.)*

As regards the process that raters follow in assigning a score, raters mostly use holistic scoring and only assign analytic feedback as an afterthought. However, it appears that the analytic criteria play an important role in helping teachers apply the rating scale to arrive at a holistic score. So, although the analytic criteria may not provide accurate or useful feedback to students, it may be necessary for raters and could contribute to inter-rater reliability. Raters are more likely to perform more consistently, both as a group and as individuals, if they base their decisions off a common set of mutually understood criteria.

In order to improve the validity of the LAT, the raters need to improve the degree to which they prioritize and value the criteria and writing samples in the portfolios. If raters greatly differ in their grading of portfolios, it could affect exam scores, and, as a result, the standard for what constitutes a 3 or a 3+ (etc.) portfolio could become confused. LAT validity will improve as there is greater mutual understanding among raters as to what represents a portfolio of any given point on the rating scale.

Recommendations for LATs

The scoring-related validity of the exam could be improved by implementing three initial changes to the writing portfolio LAT. First, rater modeling process (which has already been done in rater training sessions) appears to help raters hone their own rating process. Moreover, this modeling appears to encourage a more uniform process among the raters which may contribute to both inter-rater and intra-rater reliability. If more rating process modeling is done throughout the semester, it may help raters both in their roles as raters as well as teachers.

Second, several raters emphasized the need for a mandatory multi-draft essay that exercised the most challenging level objectives. This could help increase both the content validity of the LAT and could help improve rater consistency by further emphasizing the requirements for benchmark portfolios. The ELC writing program coordinator may also wish to re-evaluate and better define the role that the 30-minute and Metacognitive essays in portfolio assessment. This would then need to be clearly explained to and understood by raters.

Third, the MRFM analysis of the writing criteria revealed that the current method of providing and measuring analytic feedback is not very effective. The results for all criteria are highly correlated to the holistic score. This could be improved if a secondary scale, that gave more precise feedback, were used for measuring just the writing criteria. This more useful feedback could help individual students target areas for particular personal difficulty. This scale could also be used in a repeat analysis of writing criteria and could indicate whether any specific criteria tend to be more challenging for ELC students and then address those student deficiencies in their teaching.

Implications for Teaching

Lack of consensus in rating criteria is not just a LAT problem, but a classroom issue as well. If teachers do not agree on which criteria is most important, students may be confused as they move from one level to another. For example, if a Level 2 teacher stresses brainstorming, drafting, and revision, then students in that class may believe that the writing process is critically important to success at the ELC. However, if the following semester they are taught by a teacher who feels that the writing process is not an important aspect of good writing, then it will likely confuse students and they may receive low grades from this teacher who feels that they are wasting time on low priority skills. They may also receive a low LAT score if the raters do not hold the same criteria priorities as their classroom teacher.

In analyzing these results, it became clear that raters differed not only on what rating criteria to prioritize, but also on what types of writing assignments were appropriate for each level. This is leading to a redefinition of writing program objectives for each level and is helping the writing program coordinator and writing teachers to select one mandatory level-appropriate essay for the LAT portfolio. For example, instructors now receive a detailed list of writing level objectives along with definitions for those terms. Then all instructors in a given level decide upon writing assignments for the course that will exercise the writing objectives for that level. One of these assignments, which encourages the use of the most challenging objectives for the given level, is selected as the level compulsory paper. This compulsory paper must be included in the writing portfolio so that there is a more common basis of skill ability when raters

grade portfolios are a given level. This should help improve the teaching of writing as well as help raters to achieve a greater consensus of LAT assessment.

Finally, the MFRM analysis (see Table 4.3) revealed that raters assign higher scores to global writing skills (organization, content, writing process, etc.) than local skills (grammar, vocabulary, etc.). This suggests that students are performing well on global factors but are weak with local issues. Students may benefit from more specific language-related writing instruction in addition to the excellent global composition skills instruction that they are currently receiving.

Limitations

This study was limited to 21 raters during a single semester at the ELC. Results from only 251 students were used, with less than 30 students in one level. The population trends for the ELC could be very different from one semester to the next in the case of students as well as raters. Consequently, these results are not generalizable to all students or raters at the ELC, and they are certainly not generalizable to other EIL programs. The conclusions about examinees performance, criteria prioritizing, and rater processes may only be applicable to the ELC's writing LAT context.

The encouraging results from the Level/Class MFRM analysis could be due in part to a predisposed separation of students into appropriate rating categories. These somewhat artificially positive results are the product of the current rating process. Raters are given a set of portfolios and are told which level, and hence which range of the rating scale they should use when assigning a score. As such, students in a particular level are always within a 5-point range on the scale. There is still room for variation within those five points, and disordering of classes, levels, or students could still occur, but in general,

the level subscales may give an artificial impression regarding the effectiveness of the LATs. This should be considered when interpreting the MRFM analysis.

Suggestions for Further Research

This study has taken a holistic look at writing exam rater processes. This approach was necessary in order to fulfill the requirements of a well-balanced validity study. However, this investigation has raised a number of questions about specific aspects of the rating process. First, it may be beneficial to investigate the relationships among scores and rater factors such as the length of rater teaching experience, the length of rater rating experience, and the breath of rater rating experience across levels. Insight into rater factors may help program administrators select the most reliable raters or gain insight into how to train raters to be more reliable.

In addition to rater factors, essay factors is another possible area for further research. As noted earlier, some raters were concerned that students may have been unfairly penalized or rewarded based on the selection of writing samples included in the portfolio. Researchers could measure the relationship that essay type has on portfolio score. Other essay factors include essay length and topic, as well as language features such as grammar, sentence structure, and vocabulary.

One of the anticipated benefits of the new rating scale for the ELC writing LAT is the expectation of linking the ELC writing program to composition courses at BYU, the sponsoring institution. All new non-native English-speaking international students at BYU must take a writing placement exam before they are placed into a writing course. Those who pass the exam may take English Language (ELANG) 105, a composition course designed for non-native English-speaking international students that serves as an

option for general education composition credit. Those who fail the placement exam are enrolled in English as a Second Language (ESL) 304, a supplementary writing course that prepares students for ELANG 105. Currently, all ELC students who are accepted to full-time study at BYU must take this placement test. However, because the LAT rating scale now allows for two graduating levels (6 and 7), these scores could be used to indicate candidacy for ESL 304, ELANG 105, or another general education composition course.

As discussed in the limitations section, under the current rating process, raters for a particular level only deal with a 5-point subsection of the scale which may lead to artificially positive results about the effectiveness of the LAT. In order to truly gauge raters' ability to distinguish among the scale categories, a study in blind ratings could be done: one set of raters could rate portfolios with the level-appropriate subscales, and a second group of raters could rate unmarked portfolios from any level using the whole scale. If the effectiveness of raters from both groups are equally good, then it would help support the results of this study. If not, then training may be needed to help raters better distinguish between portfolios of varying proficiency.

Conclusion

This study has shown how qualitative and quantitative research methods can be combined to give a more detailed inquiry into the scoring-related validity of a writing exam. Additionally, this study also demonstrates that how the quantitative analysis (MFRM analysis) and qualitative (rater surveys, rater TAPs, and rater interviews) can be applied to a multi-level portfolio ESL program. The findings of this study are not intended to be generalizable to a larger population. Rather, this research serves as an

example inquiry into a specific testing situation. It is the process and tools, more so than the particular results, of this research that make it a valuable contribution to the field of language testing. The combination of quantitative and qualitative analyses gives a more rounded view of the testing situation, helps interpret results in a more complete way, and can be used to validate aspects of a particular exam. Just as the results of this study have influenced the writing program at the ELC, program administrators at other institutions can then use the results from their own validity study to make improvements to their own writing portfolio exam.

References

Arkoudis. S., & O'Loughlin, K. (2004). Tensions between validity and outcomes: Teacher assessment of written work of recently arrived immigrant ESL students. *Language Testing, 21*, 284-304.

Bachman, L.F. (2002a). Alternate interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practices, 21*, 5-17.

Bachman, L.F. (2002b). Some reflections on task-based language performance assessment. *Language Testing, 19*, 435-476.

Bachman, L.F., Lynch, B.K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*, 238-257.

Bailey, K.M. (1998). *Learning about language assessment: Dilemmas, decisions, and directions.* New York: Heinle & Heinle.

Brown, J.D., & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge: Cambridge UP.

Brown, J.D., Hudson, T., Norris, J. & Bonk, W.J. (2002). *An investigation of second language task-based performance assessment.* Honolulu: U of Hawai'i P.

Campbell, C. (1998). *Teaching second-language writing: Interacting with text*. New York: Heinle & Heinle.

Cho, Y. (2003). Assessing writing: Are we bound by only one method? *Assessing Writing, 8*, 165-191.

Coombe, C. & Barlow, L. (2004). The reflective portfolio: Two case studies from the United Arab Emirates. *English Teaching Forum, 42*, 18-23.

Cumming, A. (2001). ESL/EFL instructors practice for writing assessment: Specific purposes or general purposes? *Language testing, 18*, 207-224.

Cumming, A. & Berwick, R. (1996). *Validation in language testing*. Bristol, PA: Multilingual Matters.

Ericsson, K. A. & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: The MIT Press.

Gass, S.M. & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hamp-Lyons, L. (1994). Applying ethical standards to portfolio assessment of writing in English as a second language. In M. Milanovic & N. Saville (Eds.), *Studies in language testing* (pp. 151-164). Cambridge: Cambridge University Press.

Hamp-Lyons, L. & Condon, W. (2000). *Assessing the portfolio: Principles for practice, theory, and research.* Cresskill, NJ: Hampton Press.

Hughes, A. (2003). *Testing for language teachers.* Cambridge: Cambridge UP.

Huot, B. (2002). *(Re)Articulating writing assessment for teaching and learning.* Logan, UT: Utah State UP.

Hyland, K. (2002). *Teaching and Researching Writing*. London: Longman.

Kane, M.T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kane, M.T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319-342.

Kim, H. (2006). Providing validity evidence for a speaking test using FACETS. *Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics, 6*, 1-37.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to raters? *Language Testing, 19*, 246-276.

Lee, H. (2005). *Validity of English listening level achievement tests at the English language center of Brigham Young University*. Unpublished master's thesis, Brigham Young University, UT.

Lynch, B.K. & McNamara, T.F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*, 158-180.

Messick, S. (1992). Validity of test interpretation and use. In *Encyclopedia of Educational Research* (Vol. 4, pp. 1487-1495). New York:Macmillan.

Moya, S. & O'Malley, J. (1994). A portfolios assessment model for ESL. *The Journal of Educational Issues of Language Minority Students, 13*, 13-36.

Neal, M. (1998). The politics and perils of portfolio grading. In F. Zak & C.C. Weaver (Eds.), *The theory and practice of grading writing: Problems and possibilities* (pp.123-137). Albany: State University of New York Press.

Park, T. (2004). An investigation of an ESL placement test of writing using multi-faceted Rasch measurement. *Teachers College, Columbia University Working Papers in TESOL and Applied Linguistics, 4*, 1-21.

Pollitt, A. & Murray, N.L. (1994). What raters really pay attention to. In M. Milanovic & N. Saville (Eds.), *Studies in language testing* (pp.74-91). Cambridge: Cambridge University Press.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22*, 1-30.

Song, B. & Bonne, A. (2002). Using portfolios to assess the writing of students: A powerful alternative. *Journal of second language writing, 11*, 49-72.

Tai, J. (2004). *Validity of English speaking level achievement tests at the English language center of Brigham Young University*. Unpublished master's thesis, Brigham Young University, UT.

Turner, C.E. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *Canadian Modern Language Review, 56*, 555-584.

Weir, C.J. (2005). *Language testing and validation: An evidence-based approach*. New York: Palgrave MacMillian.

APPENDIX A
Rater Feedback Sheets

Level 1 Writing Portfolio Feedback

Student: _____ Classroom Teacher: _____ Rater_____

Dear Student: The purpose of this paper is to tell you about you writing portfolio—the final exam for your writing class. Two teachers have read your portfolio and filled out feedback sheets like this one. So, you should get at least two of these feedback sheets. Together, these feedback sheets will tell you what score your writing portfolio received. They will also help you understand the strengths and weaknesses in your writing.

Score Key:

1: Needs to repeat Level 1	**1+**: Will struggle in Level 2
2: Ready to begin Level 2	**2+**: Will do very well in Level 2
3: Possibly ready to begin Level 3	

Writing Features:

Topics: **1** **1+** **2** **2+** **3**
- Your topics are at a high level.

Content: **1** **1+** **2** **2+** **3**
- Your writing contains a lot specific and clever ideas.
- Your writing shows excellent thinking and effort.

Organization: **1** **1+** **2** **2+** **3**
- You have a clear purpose for writing.
- You connect your ideas well.
- Your ideas are in a good, logical order.

Vocabulary: **1** **1+** **2** **2+** **3**
- You use a lot of different words.
- You use a lot of high level words.

Grammar: **1** **1+** **2** **2+** **3**
- Your grammar has almost no errors.
- You use high level grammar.

Editing: **1** **1+** **2** **2+** **3**
- Your spelling is correct.
- Your punctuation is correct.
- Your formatting is correct.

Writing Process: **1** **1+** **2** **2+** **3**
- You changed the content and organization in your drafts and not only grammar, spelling, and punctuation.

30-minute Essay: **1** **1+** **2** **2+** **3**
- You have a lot of great ideas.
- You organized your ideas well.

Thinking About Your Writing: **1** **1+** **2** **2+** **3**
- You understand the role of audience in your writing.
- You understand the purpose of your writing.
- You understand the writing process.

Other Comments: _____

Explanation of scores: You will get **two scores**. Use the table below to find your grade for your writing portfolio.

Portfolio Grades	
2 honors ratings = 100%	1 pass, 1 low pass = 75%
1 honors, 1 high pass = 96%	2 low pass ratings = 70%
2 high pass ratings = 93%	1 low pass, 1 no pass = 60%
1 high pass, 1 pass = 85%	2 no pass ratings = 50%
2 pass ratings = 80%	no pass for missing drafts or proven plagiarism = 0%

Portfolio Score:
This is your score from ONE rater:
- 3 (Honors)
- 2+ (High Pass)
- 2 (Pass)
- 1+ (Low Pass)
- 1 (No Pass)
- No Pass because of missing drafts
- No Pass because of proven plagiarism

Level 2 Writing Portfolio Feedback

Student: _____ Classroom Teacher: _____ Rater_____

Dear Student: The purpose of this paper is to tell you about you writing portfolio—the final exam for your writing class. Two teachers have read your portfolio and filled out feedback sheets like this one. So, you should get at least two of these feedback sheets. Together, these feedback sheets will tell you what score your writing portfolio received. They will also help you understand the strengths and weaknesses in your writing.

Score Key:

2: Needs to repeat Level 2	**2+**: Will struggle in Level 3
3: Ready to begin Level 3	**3+**: Will do very well in Level 3
4: Possibly ready to begin Level 4	

Writing Features:

Topics: **2** **2+** **3** **3+** **4**

- Your topics are at a high level.

Content: **2** **2+** **3** **3+** **4**

- Your writing contains a lot specific and clever ideas.
- Your writing shows excellent thinking and effort.

Organization: **2** **2+** **3** **3+** **4**

- You have a clear purpose for writing.
- You connect your ideas well.
- Your ideas are in a good, logical order.

Vocabulary: **2** **2+** **3** **3+** **4**

- You use a lot of different words.
- You use a lot of high level words.

Grammar: **2** **2+** **3** **3+** **4**

- Your grammar has almost no errors.
- You use high level grammar.

Editing: **2** **2+** **3** **3+** **4**

- Your spelling is correct.
- Your punctuation is correct.
- Your formatting is correct.

Writing Process: **2** **2+** **3** **3+** **4**

- You changed the content and organization in your drafts and not only grammar, spelling, and punctuation.

30-minute Essay: **2** **2+** **3** **3+** **4**

- You have a lot of great ideas.
- You organized your ideas well.

Thinking About Your Writing: **2** **2+** **3** **3+** **4**

- You understand the role of audience in your writing.
- You understand the purpose of your writing.
- You understand the writing process.

Other Comments: _____

Explanation of scores: You will get **two scores**. Use the table below to find your grade for your writing portfolio.

Portfolio Grades	
2 honors ratings = 100%	1 pass, 1 low pass = 75%
1 honors, 1 high pass = 96%	2 low pass ratings = 70%
2 high pass ratings = 93%	1 low pass, 1 no pass = 60%
1 high pass, 1 pass = 85%	2 no pass ratings = 50%
2 pass ratings = 80%	no pass for missing drafts or proven plagiarism = 0%

Portfolio Score:
This is your score from ONE rater:

4 (Honors)
3+ (High Pass)
3 (Pass)
2+ (Low Pass)
2 (No Pass)
No Pass because of missing drafts
No Pass because of proven plagiarism

Level 3 Writing Portfolio Feedback

Student: _____ Classroom Teacher: _____ Rater_____

Dear Student: The purpose of this paper is to tell you about you writing portfolio—the final exam for your writing class. Two teachers have read your portfolio and filled out feedback sheets like this one. So, you should get at least two of these feedback sheets. Together, these feedback sheets will tell you what score your writing portfolio received. They will also help you understand the strengths and weaknesses in your writing.

Score Key:

3: Needs to repeat Level 3	**3+**: Will struggle in Level 4
4: Ready to begin Level 4	**4+**: Will do very well in Level 4
5: Possibly ready to begin Level 5	

Writing Features:

Topics: **3** **3+** **4** **4+** **5**

- The topics of your papers were complex and challenging.

Content: **3** **3+** **4** **4+** **5**

- Your writing contains a lot of concrete and sophisticated ideas.
- Your writing shows excellent thinking and effort.

Organization: **3** **3+** **4** **4+** **5**

- You have a clear purpose for writing.
- You use transitions appropriately.
- You put your paragraphs in a logical order.
- You put the ideas in each paragraph in a logical order.

Vocabulary: **3** **3+** **4** **4+** **5**

- You use a good variety of words.
- You use academic words in your writing.

Grammar: **3** **3+** **4** **4+** **5**

- Your grammar has almost no errors.
- You use difficult and complex grammar.

Editing: **3** **3+** **4** **4+** **5**

- Your spelling is correct.
- Your punctuation is correct.
- Your formatting is correct.

Writing Process: **3** **3+** **4** **4+** **5**

- You focused on revising content and organization throughout your drafts and not only grammar, spelling, and punctuation.

30-minute Essay: **3** **3+** **4** **4+** **5**

- You have a lot of great ideas.
- You organized your ideas well.

Thinking About Your Writing: **3** **3+** **4** **4+** **5**

- You understand the role of audience in your writing.
- You understand the purpose of your writing.
- You understand the writing process.

Other Comments: _____

Explanation of scores: You will get **two scores**. Use the table below to find your grade for your writing portfolio.

Portfolio Grades	
2 honors ratings = 100%	1 pass, 1 low pass = 75%
1 honors, 1 high pass = 96%	2 low pass ratings = 70%
2 high pass ratings = 93%	1 low pass, 1 no pass = 60%
1 high pass, 1 pass = 85%	2 no pass ratings = 50%
2 pass ratings = 80%	no pass for missing drafts or proven plagiarism = 0%

Portfolio Score:
This is your score from ONE rater:

5 (Honors)
4+ (High Pass)
4 (Pass)
3+ (Low Pass)
3 (No Pass)
No Pass because of missing drafts
No Pass because of proven plagiarism

Level 4 Writing Portfolio Feedback

Student: _____ Classroom Teacher: _____ Rater_____

Dear Student: The purpose of this paper is to tell you about you writing portfolio—the final exam for your writing class. Two teachers have read your portfolio and filled out feedback sheets like this one. So, you should get at least two of these feedback sheets. Together, these feedback sheets will tell you what score your writing portfolio received. They will also help you understand the strengths and weaknesses in your writing.

Score Key:

4: Needs to repeat Level 4	4+: Will struggle in Level 5
5: Ready to begin Level 5	5+: Will do very well in Level 5
6: Possibly ready to begin BYU writing classes	

Writing Features:

Topics: **4 4+ 5 5+ 6**

- The topics of your papers were complex and challenging.

Content: **4 4+ 5 5+ 6**

- Your writing contains a lot of concrete and sophisticated ideas.
- Your writing shows excellent thinking and effort.

Organization: **4 4+ 5 5+ 6**

- You have a clear purpose for writing.
- You use transitions appropriately.
- You put your paragraphs in a logical order.
- You put the ideas in each paragraph in a logical order.

Vocabulary: **4 4+ 5 5+ 6**

- You use a good variety of words.
- You use academic words in your writing.

Grammar: **4 4+ 5 5+ 6**

- Your grammar has almost no errors.
- You use difficult and complex grammar.

Editing: **4 4+ 5 5+ 6**

- Your spelling is correct.
- Your punctuation is correct.
- Your formatting is correct.

Writing Process: **4 4+ 5 5+ 6**

- You focused on revising content and organization throughout your drafts and not only grammar, spelling, and punctuation.

30-minute Essay: **4 4+ 5 5+ 6**

- You have a lot of great ideas.
- You organized your ideas well.

Thinking About Your Writing: **4 4+ 5 5+ 6**

- You understand the role of audience in your writing.
- You understand the purpose of your writing.
- You understand the writing process.

Other Comments: _____

Explanation of scores: You will get **two scores**. Use the table below to find your grade for your writing portfolio.

Portfolio Grades	
2 honors ratings = 100%	1 pass, 1 low pass = 75%
1 honors, 1 high pass = 96%	2 low pass ratings = 70%
2 high pass ratings = 93%	1 low pass, 1 no pass = 60%
1 high pass, 1 pass = 85%	2 no pass ratings = 50%
2 pass ratings = 80%	no pass for missing drafts or proven plagiarism = 0%

Portfolio Score:
This is your score from ONE rater:

6 (Honors)
5+ (High Pass)
5 (Pass)
4+ (Low Pass)
4 (No Pass)
No Pass because of missing drafts
No Pass because of proven plagiarism

Level 5 Writing Portfolio Feedback

Student: _____ Classroom Teacher: _____ Rater_____

Dear Student: The purpose of this paper is to tell you about you writing portfolio—the final exam for your writing class. Two teachers have read your portfolio and filled out feedback sheets like this one. So, you should get at least two of these feedback sheets. Together, these feedback sheets will tell you what score your writing portfolio received. They will also help you understand the strengths and weaknesses in your writing.

Score Key:

5: Ready to begin Level 5	**5+**: Will struggle in BYU classes
6: Possibly ready to begin BYU classes	**6+**: Possibly will do very well in BYU classes
7: Possibly will do extremely well in BYU classes	

Writing Features:

Topics: **5** **5+** **6** **6+** **7**

- The topics of your papers were complex and challenging.

Content: **5** **5+** **6** **6+** **7**

- Your writing contains a lot of concrete and sophisticated ideas.
- Your writing shows excellent thinking and effort.

Organization: **5** **5+** **6** **6+** **7**

- You have a clear purpose for writing.
- You use transitions appropriately.
- You put your paragraphs in a logical order.
- You put the ideas in each paragraph in a logical order.

Vocabulary: **5** **5+** **6** **6+** **7**

- You use a good variety of words.
- You use academic words in your writing.

Grammar: **5** **5+** **6** **6+** **7**

- Your grammar has almost no errors.
- You use difficult and complex grammar.

Editing: **5** **5+** **6** **6+** **7**

- Your spelling is correct.
- Your punctuation is correct.
- Your formatting is correct.

Writing Process: **5** **5+** **6** **6+** **7**

- You focused on revising content and organization throughout your drafts and not only grammar, spelling, and punctuation.

30-minute Essay: **5** **5+** **6** **6+** **7**

- You have a lot of great ideas.
- You organized your ideas well.

Thinking About Your Writing: **5** **5+** **6** **6+** **7**

- You understand the role of audience in your writing.
- You understand the purpose of your writing.
- You understand the writing process.

Other Comments: _____

Explanation of scores: You will get **two scores**. Use the table below to find your grade for your writing portfolio.

Portfolio Grades	
2 honors ratings = 100%	1 pass, 1 low pass = 75%
1 honors, 1 high pass = 96%	2 low pass ratings = 70%
2 high pass ratings = 93%	1 low pass, 1 no pass = 60%
1 high pass, 1 pass = 85%	2 no pass ratings = 50%
2 pass ratings = 80%	no pass for missing drafts or proven plagiarism = 0%

Portfolio Score:
This is your score from ONE rater:

7 (Honors)
6+ (High Pass)
6 (Pass)
5+ (Low Pass)
5 (No Pass)
No Pass because of missing drafts
No Pass because of proven plagiarism

APPENDIX B

Rater Survey

ELC Writing LAT Rater Survey Fall 2005

Dear writing portfolio raters,

In our on-going efforts to improve the LATs, we ask you to please provide your feedback **regarding your mindset when rating the portfolios this semester**. You do not need to include your name on this form, just the level of portfolios you *rated*. If you rated more than one level, please fill-out a separate form for each level.

We thank you in advance for your feedback.

Here are the rater survey questions. Please complete **both front and back** of this form.

1. What is your rater number? _____

2. What level did you rate? (Please circle one only one level per survey.)

1 2 3 4 5

3. As you were rating writing portfolios this semester, what writing features did you place the greatest emphasis on when assigning a score? Please rank the following writing criteria by circling a number from not at all important (1) to very important (5):

A. topic difficulty

Not at all important — Somewhat important — Very important

1 2 3 4 5

B. interesting content

Not at all important — Somewhat important — Very important

1 2 3 4 5

C. length of papers

Not at all important — Somewhat important — Very important

1 2 3 4 5

D. depth of topic

Not at all important — Somewhat important — Very important

1 2 3 4 5

E. organization and order

Not at all important — Somewhat important — Very important

1 2 3 4 5

F. depth/variety of grammar usage

Not at all important — Somewhat important — Very important

1 2 3 4 5

G. accuracy of grammar usage

Not at all important		Somewhat important		Very important
1	2	3	4	5

H. vocabulary

Not at all important		Somewhat important		Very important
1	2	3	4	5

I. spelling

Not at all important		Somewhat important		Very important
1	2	3	4	5

J. formatting

Not at all important		Somewhat important		Very important
1	2	3	4	5

K. punctuation

Not at all important		Somewhat important		Very important
1	2	3	4	5

L. writing process and drafts

Not at all important		Somewhat important		Very important
1	2	3	4	5

M. 30 minute writing sample

Not at all important		Somewhat important		Very important
1	2	3	4	5

N. metacognitive essay

Not at all important		Somewhat important		Very important
1	2	3	4	5

If you rate portfolios based on additional criteria not included above, please indicate so.

O. other: _____

Not at all important		Somewhat important		Very important
1	2	3	4	5

P. other: _____

Not at all important		Somewhat important		Very important
1	2	3	4	5

Q. other: _____

Not at all important		Somewhat important		Very important
1	2	3	4	5

APPENDIX C

MFRM Measurement Report for Levels

Level ID	Ability (logits)	Standard Error	Infit MS
L5	+2.91	0.06	1.02
L4	+1.46	0.04	0.85
L3	- 0.12	0.04	0.82
L2	- 1.49	0.05	0.83
L1	- 2.75	0.07	0.80
Mean	0.00	0.09	0.86
Standard Deviation	2.26	0.02	0.09

Reliability of separation index = 1.00

APPENDIX D

MFRM Measurement Report for Classes

Class ID	Ability (logits)	Standard Error	Infit MS
5A	+3.94	0.09	1.04
5B	+3.53	0.09	1.04
5C	+3.20	0.12	0.94
4A	+1.91	0.10	0.80
4C	+1.85	0.10	0.62
4D	+1.67	0.07	0.95
4B	+1.59	0.07	0.89
3D	+0.73	0.09	0.73
3C	+0.63	0.07	0.92
3A	+0.31	0.09	0.96
3B	+0.24	0.06	0.70
2B	- 0.47	0.09	0.83
2A	- 0.91	0.09	0.69
2C	- 1.10	0.09	0.97
1A	- 2.09	0.08	0.88
1B	- 2.42	0.14	0.58
Mean	0.79	0.09	0.85
Standard Deviation	1.90	0.02	0.14

Reliability of separation index = 1.00

APPENDIX E

MFRM Measurement Report for Examinees

Examinee ID	Ability (logits)	Standard Error	Infit MS
501413	5.91	0.33	0.69
501310	5.18	0.32	1.13
501303	5.12	0.27	1.48
501309	4.77	0.26	0.65
501503	4.57	0.32	1.04
501509	4.34	0.33	1.06
501410	4.22	0.33	0.43
401201	4.16	0.36	0.63
401218	3.93	0.36	0.50
501306	3.90	0.33	0.24
501402	3.82	0.32	0.58
401012	3.77	0.36	0.13
501305	3.69	0.33	0.36
501312	3.66	0.27	1.22
401014	3.64	0.36	0.20
401109	3.64	0.36	0.08
501408	3.54	0.34	0.73
501311	3.51	0.27	1.53
501406	3.43	0.34	0.66
501506	3.30	0.34	0.66
401204	3.25	0.36	0.95
501411	3.21	0.34	1.39
400905	3.16	0.36	0.71
401011	3.10	0.36	0.41
501508	2.95	0.35	1.11
501407	2.85	0.35	0.48
501302	2.76	0.34	1.74
400901	2.75	0.36	0.48
501404	2.71	0.34	1.27
501304	2.64	0.34	1.03
501409	2.61	0.35	1.65
300802	2.48	0.34	0.47
501403	2.47	0.34	0.58
501507	2.45	0.36	1.08
401225	2.35	0.35	1.04
401010	2.34	0.35	0.59
401110	2.32	0.35	0.80
501301	2.30	0.34	2.04
400908	2.29	0.35	0.96
401113	2.29	0.35	0.59
501308	2.20	0.35	0.39
300606	2.12	0.28	0.78
501502	2.08	0.35	1.54
501401	1.99	0.35	2.36
301603	1.97	0.34	0.80
401106	1.96	0.35	0.60
300709	1.89	0.34	1.70
401220	1.89	0.28	3.26
401015	1.88	0.35	1.07
501412	1.86	0.36	2.01

Examinee ID	Ability (logits)	Standard Error	Infit MS
401102	1.83	0.35	0.34
501307	1.83	0.35	0.38
501501	1.83	0.36	0.96
300712	1.76	0.34	0.56
301613	1.74	0.34	0.37
400902	1.65	0.35	1.02
400904	1.65	0.35	1.10
400911	1.65	0.35	1.48
401219	1.65	0.35	1.29
401018	1.61	0.35	1.60
301607	1.53	0.34	0.49
401211	1.53	0.35	0.43
401223	1.50	0.35	1.06
300801	1.44	0.34	0.31
300803	1.44	0.34	0.87
300812	1.43	0.34	0.82
301604	1.38	0.34	1.32
300607	1.33	0.35	0.74
401016	1.25	0.35	0.81
300610	1.21	0.35	0.51
300723	1.19	0.34	1.45
300810	1.17	0.34	0.47
401003	1.15	0.34	0.34
401224	1.15	0.34	0.37
200302	1.11	0.33	0.63
400909	1.08	0.35	0.59
401210	1.05	0.35	0.37
401017	1.02	0.34	2.37
401103	1.00	0.34	0.29
401203	0.95	0.34	0.44
401212	0.93	0.34	0.69
401217	0.93	0.28	1.85
401002	0.91	0.34	0.57
300808	0.86	0.34	0.24
401214	0.85	0.34	0.94
401216	0.85	0.34	0.27
401013	0.82	0.34	0.70
401205	0.82	0.34	1.11
400912	0.79	0.34	0.42
400913	0.79	0.34	0.48
300611	0.78	0.35	0.58
401006	0.78	0.28	1.77
300625	0.74	0.35	0.47
401112	0.73	0.34	0.30
300703	0.72	0.34	1.41
300706	0.72	0.34	0.58
300708	0.72	0.34	0.68
300717	0.72	0.34	2.18
401202	0.72	0.34	0.43
501405	0.72	0.36	0.97
400910	0.70	0.34	0.98
401004	0.67	0.34	1.08
501504	0.65	0.36	0.67
401215	0.62	0.34	0.37
300705	0.60	0.35	0.55

Examinee ID	Ability (logits)	Standard Error	Infit MS
300713	0.58	0.35	1.33
401207	0.58	0.34	0.91
401222	0.56	0.34	0.83
401101	0.54	0.34	0.57
300722	0.47	0.35	0.72
300601	0.41	0.35	1.14
300618	0.40	0.35	0.93
501505	0.40	0.36	0.30
300715	0.39	0.35	0.47
300724	0.36	0.35	1.03
200403	0.34	0.33	0.44
401001	0.34	0.34	0.46
300623	0.14	0.35	0.61
300701	0.14	0.35	0.55
301610	0.06	0.35	1.43
401111	0.06	0.34	1.04
401107	0.05	0.34	0.59
300702	0.02	0.35	0.69
401209	0.00	0.34	0.68
401009	-0.03	0.34	0.77
401108	-0.14	0.34	0.59
301602	-0.15	0.35	0.45
401008	-0.15	0.34	0.72
401019	-0.17	0.34	0.56
401020	-0.17	0.34	0.55
300605	-0.21	0.35	0.42
300626	-0.22	0.35	0.88
300813	-0.22	0.35	0.70
300619	-0.23	0.35	0.37
401206	-0.23	0.34	1.52
401221	-0.25	0.34	0.62
300704	-0.34	0.35	1.31
300718	-0.34	0.35	1.00
300721	-0.38	0.35	0.43
400903	-0.45	0.34	0.83
300814	-0.46	0.35	0.92
300622	-0.47	0.34	0.56
200412	-0.48	0.33	1.50
300710	-0.50	0.35	0.77
300711	-0.50	0.35	1.32
400906	-0.59	0.34	0.92
300707	-0.60	0.35	0.43
401005	-0.60	0.34	1.59
100112	-0.63	0.33	0.62
200406	-0.63	0.33	0.17
300604	-0.68	0.35	0.54
400907	-0.70	0.34	0.45
200306	-0.72	0.33	0.28
401007	-0.72	0.34	1.48
200510	-0.75	0.33	0.70
301612	-0.76	0.34	1.29
401213	-0.77	0.34	0.64
301611	-0.78	0.34	1.20
200413	-0.80	0.33	1.03
300804	-0.84	0.34	0.84

Examinee ID	Ability (logits)	Standard Error	Infit MS
300805	-0.84	0.34	1.31
301601	-0.87	0.28	1.56
200313	-0.90	0.33	0.44
300603	-0.92	0.35	0.75
300620	-0.93	0.34	0.84
200401	-0.94	0.33	1.64
401105	-0.98	0.34	1.61
200312	-1.01	0.33	0.75
300613	-1.01	0.34	0.68
300807	-1.02	0.35	0.25
200501	-1.07	0.33	0.74
200309	-1.15	0.33	1.00
200402	-1.15	0.33	0.44
300608	-1.15	0.33	0.69
300621	-1.15	0.33	0.94
100106	-1.18	0.33	1.15
200508	-1.18	0.33	0.61
401021	-1.21	0.34	0.20
200404	-1.26	0.33	0.39
300714	-1.28	0.35	0.36
300809	-1.28	0.35	1.41
200503	-1.29	0.33	0.72
301606	-1.32	0.34	0.86
301608	-1.35	0.34	0.71
300806	-1.38	0.35	0.69
100117	-1.39	0.33	0.36
200405	-1.43	0.27	1.67
300720	-1.54	0.33	0.73
200408	-1.59	0.33	0.30
100103	-1.67	0.27	1.29
200414	-1.68	0.34	0.79
100102	-1.72	0.33	0.93
300602	-1.73	0.34	0.96
200502	-1.74	0.34	1.08
300716	-1.75	0.34	0.66
300811	-1.75	0.33	0.91
301614	-1.77	0.33	0.96
300719	-1.78	0.34	1.23
300627	-1.80	0.33	0.70
300612	-1.81	0.33	0.86
300617	-1.84	0.33	0.70
301605	-1.87	0.33	1.19
300624	-1.90	0.33	0.65
200310	-1.92	0.34	0.19
300614	-2.03	0.33	0.58
200303	-2.04	0.34	0.86
200305	-2.04	0.34	0.68
200512	-2.07	0.34	0.84
200507	-2.08	0.34	0.42
300609	-2.11	0.33	0.46
200411	-2.13	0.34	0.82
301609	-2.14	0.33	0.58
200307	-2.16	0.34	0.44
100107	-2.19	0.28	0.90
100104	-2.28	0.34	0.33

Examinee ID	Ability (logits)	Standard Error	Infit MS
200409	-2.37	0.34	0.71
300615	-2.39	0.33	1.30
200511	-2.40	0.33	0.35
200513	-2.40	0.33	0.68
200506	-2.42	0.34	0.87
100206	-2.55	0.36	0.44
200410	-2.58	0.34	0.34
100207	-2.68	0.36	0.45
200311	-2.70	0.33	0.57
200314	-2.70	0.33	0.85
200407	-2.71	0.34	0.91
200504	-2.75	0.33	3.61
200509	-2.75	0.33	0.76
100204	-2.81	0.37	0.42
300616	-2.81	0.33	0.30
200304	-2.94	0.33	1.07
100202	-2.95	0.38	0.83
100111	-3.01	0.36	1.14
100203	-3.10	0.39	0.65
100109	-3.15	0.37	0.42
200308	-3.16	0.33	1.27
200514	-3.16	0.27	1.09
200505	-3.19	0.33	1.07
100201	-3.28	0.40	0.54
200301	-3.59	0.33	0.66
100108	-3.75	0.41	1.10
100205	-3.76	0.43	0.44
100114	-3.92	0.42	0.77
100105	-4.10	0.43	0.41
100110	-4.10	0.43	0.44
100208	-4.15	0.44	0.99
100118	-4.48	0.44	0.28
100101	-5.26	0.42	0.74
100115	-5.43	0.41	0.97
100116	-5.75	0.38	1.78
100113	-6.57	0.31	1.42
Mean	0.00	0.34	0.84
Standard Deviation	2.20	0.02	0.49

Reliability of separation index = 0.96

APPENDIX F

Rater Think-Aloud Protocol Transcripts

Rater 12 – Rating Level 1

Essay 1

I wish this paper were typed, but since this is Level 1, I won't be paying attention to that.

Well, first I think the student understands the topic. It's about a hero or king or something of a country.

But the paper is not well formatted, because at the beginning it should have been indented, every paragraph.

And he's got a nice introduction and some topic sentences.

It would be great if the student put more details in it.

And there are some places he should have used past tense instead of present tense because it's about a president's life. And there are some missing verbs.

And I think his teacher gives him comments, some good comments on the second draft, and he made those changes. So that's good.

Essay 2

And both topics are Level 1 [appropriate] topics. And one is more difficult than the other. And I think he handled both topics pretty well.

My Beautiful Family, this paper, it doesn't have a conclusion, so that's something missing there.

And I think he's writing some repeated grammatical mistakes. Like "I am choose" and "My father very young people." Some missing helping words or parts.

Metacognitive Essay

And he has a good understanding of the audience. And he mentioned that he changed the writing style because the audience couldn't understand his essay. Not very specific, but he's got something in there.

And it seems like he understands the organization as well.

And he also mentions some changes he made based on the comments made by the teacher and the classmates.

30 minute Essay

I think the 30 minute essay reflects the skills he showed in other papers and in Thinking About My Essay [metacogitive].

He didn't finish the essay, but I don't think that's important. If he had had more time, he would have finished it. But that's not something I look at.

The 30 minute essay is well organized. It's got an introduction, three body paragraphs, and a conclusion. Well, some topic sentences are not really sentences, but they are just phrases.

Overall

And grammar is still a concern for this Level 1 student. So overall, I think I will give it a "Pass" [score of 2 on the rating scale]… yeah.

[Total time 7:23]

Rater 13 – Rating Level 2

Okay, the first thing I notice is that the topics are “Joseph Smith” and “The Death Penalty.” So I can see that they are going to be a level appropriate topic.

Essay 1

This introduction: I don’t see a thesis statement.

Umm, there’s some run-on sentences.

I’m noticing – because of lack of thesis statement – the organization is not as solid or concrete as it could be, and it’s kind of jumpy from paragraph to paragraph.

The grammar in this paper is really good. It has well structured sentences. The content is a bit lacking. It has a good topic, but I don’t see much of a focus. And the conclusion kind of goes off topic and doesn’t really tie back into what it’s trying to tell me. Umm, the vocabulary is not bad.

Looking at the writing process… looks like this person just changed a lot of grammar errors.

It just goes though the life of Joseph Smith and doesn’t have much substance to it. I’m not even sure why they wrote it.

Essay 2

Going to read the second essay…

The first thing I notice is that the introduction is going from general to specific, which is a good writing technique, and they do finish the introduction with a thesis statement. But the thesis statement is… it’s in fragmented sentences.

I’m seeing a lot of good English phrases, and good use of language, good grammatical structures. Umm, however, there is a word that says “reprehend.” “Society needs rules to reprehend these crimes.” I’m not sure what that means.

I also see that the author is using sources to help support their ideas. The sources are used rather well which is pretty good for Level 2. And I see that the author has a lot of good ideas… but the grammar does get in the way towards the end.

It looks like they added a lot to the ending of the essay, which is probably why it didn’t get checked [by a tutor or teacher]. Considering correctness – I’m talking about the grammar. But because of the topic content, the topic does come through.

As of right now, I feel that it is a pass [score of 3], I'm looking at a pass. I'm going to look at the metacogintive and 30 minute [essays] to confirm that. And see how well they organize their ideas here.

Metacognitive essay

This person does understand the writing process and their audience rather well. They even say that after he went to see the tutor, he asked members of the class who were members of the Church, to see if he didn't make those changes, then they couldn't understand. He says that he thinks the paper is good no matter what, but so… I think that's rather good that he supports his writing. Although his grammar is getting in the way of a lot.

30 minute essay

Just from a glance, he does organize his essay into five paragraphs.

He does have a thesis statement, which is good. He does support each paragraph. And does tie it back to his thesis statement.

Overall

I'm going to continue with the pass score [3].

Topics: a 3.

Content: a 3.

Organization: a 3+.

Vocabulary: a 3.

Grammar: a 2+.

Editing: a 3.

The writing process: I give it a 3+.

30 minute: a 3. And thinking about your writing: a 3.

[Total time 14:40]

Rater 14 – Rating Level 3

Essay 1

Good thesis.

Nice word.

Very organized.

Let's check to see if they followed teacher's comments in improving their drafts… yes… good… this one [draft] looks good.

Essay 2

Nice complex idea there.

Good thesis.

Nice word.

Very organized.

Now let's check to see if they followed the teacher's advice for revision… yes… good.

Both papers have really good topics, complex ideas, and good grammar.

Metacognitive essay

Audience here.

Process....? Included.

Let's see; do they have purpose? Yes… here it is.

30 minute essay

Nice thesis for a 30 minute essay.

Overall

Okay, let's go to the sheets and figure out the final grade.

Topics: "Were complex and challenging?" Yes. They were. Deserves a 4+ [high pass].

"Your writing contained a lot of complex ideas." Also a 4+.

Clear purpose? Very organized… and transitions? Great. “Paragraphs in logical order?” definitely. Between a 4 and a 4+.

“Used a variety of words?” Definitely.

Grammar? No errors. Al of these [criteria] getting 4+s. One with a 4.

Spelling is correct? Yes. 4+.

Writing process…?

The metacognitive was not as clear. Deserves a 4, but was very good.

Nice organization and good ideas for a 30 minute essay.

I think that it deserves a 4.

Okay, this was definitely a very strong portfolio. Not quite an Honors [score of 5], but definitely a high pass [score of 4+] because of its diversity, clarity, and organization.

[Total time 9:32]

Rater 15 – Rating Level 3

Essay 1

The first paper here is a very high level topic. It's about solving crime in the student's home country. It provides some background knowledge and gives the points that are in the thesis statement.

She tries to build her paragraphs bit by bit. That's her paragraph organization. But some of them [points] don't really belong. It's another topic.

The conclusion is a little weak, but all the parts are there.

In the drafts, looks like she made some suggestions. Mostly it was grammar or microlinguistic feedback. She didn't make many big changes, but it look like she didn't get any feedback on that.

Essay 2

The second paper – the topic kind of disturbs me because it's just about "My Family" which is a really low level topic. And this is Level 3. This is a topic that would be appropriate for Level 1, so it's really not a good choice for Level 3.

She has good organization and lots of specific examples. It's well formatted and well organized. Lots of details which makes it a long essay.

Even though it's just an essay describing her family and her life, she does some compare and contrast – good Level 3 skills.

And there's a sentence here that shows that this is probably one of the first essays written in this semester, because the student makes reference to a recent activity [at the time that she write the essay] here at the English Language Center that was the first month of school. It helps me understand why the topic is easy. It could have been that the teacher chose the topic to help ease the students into the semester. I would not have suggested that the student include this easy topic paper in their portfolio, but I think that it helps me understand why such a simple topic was chosen. I tend to be more lenient on a paper if I know it came from earlier in the semester. Ones that I know come from the end of the semester, including the 30 minute essay, I tend to place more emphasis on, because I know that that's where the student is now.

Metacognitive essay

There's a big problem with run-on sentences here.

I can see there's some organization.

And here she makes mention that the crime paper was her last essay. Which is a more difficult topic. And she didn't have enough details. Her organization and ideas weren't as solid as that first paper. I think that is an indication that the more difficult topic, the more difficult it is to communicate that [organization and ideas].

30 minute essay

And again, big run-on sentence problem here.

There's an attempt to understand the counter-argument which is good.

Overall

I'm going to look at the feedback sheet here.

Topics: were not so complex, actually. I'm going to put that down to a 3+ [low pass]. One of them was complex; the other was way below level. I think it balances out at about a 3+.

She does use specific ideas, in particular in the one essay.

She has good organization.

Her vocabulary is… just average.

Her grammar is not so good; her editing is not very good either. Those are both 3+.

She has a good understanding of the writing process.

The 30 minute essay is just average, and her metacognitive is about average, I think.

So I'm giving her overall a pass [score of 4].

[Total time 7:03]

Rater 17 – Rating Level 4

30 minute essay

I like to start with the 30 minute because then you get an idea what their writing is like on their own. Although this time I think the question is a little bit lame.

The first thing I notice is that they don't have paragraphs; already that hurts the organization.

This first sentence is confusing.

[Student] changes the question a bit too, to children and zoos. (The topic was to agree/disagree with the effectiveness of having a zoo.)

Alright. The 30 minute essay looks confusing, although the grammar is pretty good, but the organization doesn't really answer the question.

Essay 1

This is a compare and contrast paper. His thesis includes his argument – and there is one, so that's good.

Okay, he cites his sources… does a really good job of citing his sources, actually.

These paragraphs are long; the information is good.

Alright. It looks pretty well organized. The thesis mostly fits the body paragraphs. Grammar and [sentence] structure are pretty good. It has a lot of references, and he is careful not to plagiarize; that's good.

I'm going to look at the teacher comments on the drafts to see if he revised, or if these problems were never addressed then it's less their fault if no one ever helped them.

They told him to look for the purpose and citations which he did.

Essay 2

This is his narrative.

The beginning is not very interesting. There's a lot of background and I'm wondering what the point of the story is… especially because I know his teacher and I know what she taught.

Okay… there's some [story] conflict.

This story is a little disturbing. Huh… the topic is very weird, and I wonder why he chose this. And it never says. The point of a narrative is descriptive writing and to use language effectively, which I don't think he did. And there's not citations in the narrative, so it's more important than this.

Looking at his drafts… there's a lot of grammar and vocabulary help. His first draft is two paragraphs long, which means he didn't finish it. And he didn't get much feedback on it.

That one [essay] isn't very impressive.

Metacognitive essay

He did his metacognitive [essay] on his compare and contrast [essay].

It talks about learning to use sources, which he did in that draft, actually.

Overall

Okay... ummm… his compare and contrast [essay] is much better than his narrative [essay]. I think… that is looks like… I think that I'll give overall a pass [score of 5], because his writing is okay, but it's not very… it's very… what's the word? "standard." He's… he's formulaic. And his narrative… it was… it just wasn't very good. And his 30 minute doesn't have much organization, although he does use language well; his sentences are logical and you can follow the story.

So, topics? Actually were not very good: a 4+.

Content? Umm… 4+.

Organization will get a 5.

He does better on the [sentence] structure part.

Vocabulary is a 5.

Grammar is a 5.

Editing is probably a 5.

Metacognitive is a 5.

His 30 minute essay: big 4+. And thinking about his writing is a 5.

[Total time 10:39]

Rater 20 – Rating Level 5

I basically look at the overall content. I like to look at the first and second drafts of the essays first, then I like to look at the final draft after that.

I like to see the overall presentation, get a feel for what it looks like. Ses how long it is, first of all. Go over all holistic points.

Then I start reading the beginning part to see how the flow is. To see if there is a clear understanding of what the student is trying to say in the beginning. And then I basically I go through it for the ideas, the content. Then I go back to the beginning and look for grammatical structures that might be apparent that might be errors that are common throughout.

And then, because this is Level 5, I assume that some of the papers are research papers, so I look at the way it is presented as far as detail in citing references. I don't go into a lot of detail as far as the actual detail of the references, just to make sure that they have done some research.

In going through the process here, of course I look at vocabulary: is it complex vocabulary? Sentence structure: are the sentences connected well or too simplistic versus more complex? Specifically at this level I look at that type of thing.

From that process, I can tell pretty much where a student is. This student receives a score of 6.

[Total time 2:20]

APPENDIX G

Rater Interview Responses

Rater 11

1. How does your LAT rating process differ from your essay grading process?

This semester I've implemented the LAT rating process in the way I grade my students' essays in both levels 1 and 3. In particular, I've incorporated the areas that the LAT focuses on and adjusted them slightly to match more with what I'm teaching and expecting my students be able to do. I also added values for each section in order to emphasize to the students which areas I believe, and hopefully, what I hope they soon will understand, are vital areas in improving their overall writing ability. I've used the LAT rating criteria to evaluate my students' in-class and 30-minute timed essays, so that they will become familiarized with how they'll be graded on the final writing LAT portfolio. I even have them apply a TOEFL-based criteria for self- and peer-reviewing their essays.

2. Do you feel that the LATs are a fair assessment of student ability? Why or why not?

By requiring each level to have an obligatory paper to include in their LAT portfolio is one step in helping it become a "fair" assessment of the students' abilities, but there's the issue of "fairness" when coming down to the metacognitive essay AND the timed-essay. I can see somewhat the value of the metacognitive essay in helping us to help students gain an understanding of the fact that writing IS a process, and to foster metacognitive thinking skills to help them become more independent thinkers and writers. On the other hand, I cannot clearly see the assessment value of the 30-minute timed essay. I see it usefulness in only one of 2 situations: (1) to prepare students in the Independent Writing section of the new iTOEFL, and (2) to serve as a "tiebreaker" when it comes down to finalizing. I know that in the level objectives that the students of each level are required to write so many words in 30 minutes, but this does not really support with our teaching students to apply the steps of the writing process in becoming better writers. I've had students in all levels apply the process and produce incredibly well-thought, well-organized, coherent essays in class, and then freeze up with test anxiety when taking a timed essay exam. For some students, their ability and skill in writing in-class and timed essays, may match up somewhat equally, but there are others where it does not. This is why while I assess the writing LATs, I look at the timed essay very last rather than the first.

3. How effective do you feel the LAT calibration meetings are? Do you have any suggestions for the improvement of these meeting?

They're helpful in refreshing our skills in helping teachers of all levels become more unified when it comes to assessing the writing LATs, and to help out the new teachers to the skill area become adjusted to the method of assessment. Even if you're not teaching/haven't taught a particular level, I think that it would be beneficial for teachers to assess levels that they've never taught or haven't taught for a while to practice during these meetings. Instead of relying only on the teacher who has had the most

experience teaching a particular level to be the "best" assessor, but allowing others to gain the practical experience and knowledge of assessing different levels would be beneficial for them in the long run (in terms of their professional careers).

4. Do you have any other suggestions for the improvement of the LATs?

The way that it has been improved for this semester is a step in the right direction, but having the 30-minute essay included to represent a "true" measure of a student's ability is questionable, in my own opinion, but who's to say whether it is a personal gripe or a logical argument. Who knows? It would be beneficial, I think, to perhaps have a final meeting after everyone has assessed the writing LATs to see the results directly for ourselves so that we can gain a better idea as to how [we are doing].

Rater 13

1. How does your LAT rating process differ from your essay grading process?

My LAT rating process was much more stressful because of the time pressure I was under. Because I felt like I had to get it done quickly, and because I didn't know some of the students I would have felt bad rushing through it and leaning toward the side of a lower score.

2. Do you feel that the LATs are a fair assessment of student ability? Why or why not?

Yes, because they show a student's ability throughout the writing process and on-the-spot writing. However, if the student isn't guided well during the process then I think the ability is skewed.

3. How effective do you feel the LAT calibration meetings are? Do you have any suggestions for the improvement of these meeting?

I feel the meetings are much better than they used to be. The updated benchmarks are a huge improvement and I think teachers are getting on the same page a lot faster.

4. Do you have any other suggestions for the improvement of the LATs?

I think things are going really well so far. Nothing comes to mind at the moment.

Rater 16

1. How does your LAT rating process differ from your essay grading process?

Actually my process doesn't differ. In order for me to rate well I follow the same process I use in rating essays in class so that my rating is fair and consistent. I follow the same thinking pattern and it makes it easier for me too.

2. Do you feel that the LATs are a fair assessment of student ability? Why or why not?

Sometimes. I say this because some of my students that did really well in class participating and writing good essays did not do so well in the LATs because they were

rated by someone else who has a different rating process. Also because these students put papers in their folders that were more academic than some other students but I think we have fixed that this semester with the mandatory paper. E.g. one of my lower students got a higher grade for the LATs than my more proficient students because he put in shorter easier papers and was rated accordingly and my other student put in his academic paper which warranted more severe grading.

3. How effective do you feel the LAT calibration meetings are? Do you have any suggestions for the improvement of these meeting?

Last semester could have been better. The lower level calibration meeting was good but the higher level not so much because they all came and chatted and did not really calibrate. I [would prefer that] all the calibration to be done well in advance before we get to the meeting so that we discuss the differences rather than just grading. [Also, I think it would help to choose] people that will rate better.

4. Do you have any other suggestions for the improvement of the LATs?

Written documentation of the procedure of how to administer and what needs to be done for the LATs. Giving everyone that is rating an opportunity to read the policy and procedure. Use proctors and other available resource to help with administration and little busy work.

Rater 17

1. How does your LAT rating process differ from your essay grading process?

I read a lot faster through the LATs because there are more essays and I'm just looking for an overall, holistic score. Also, I don't mark the LATs at all, so that goes faster. Oh, and because I may not know the student and it's a summative evaluation (rather than the more formative ones during the semester), I tend to grade them more formally.

2. Do you feel that the LATs are a fair assessment of student ability? Why or why not?

I think they can be, but you'd have to define what we are assessing. LATs assess their ability to write a multiple-draft essay, and they can receive help from many sources. Their 30-minute and metacognitives sort of show their writing ability without help, but it's probably not quite enough. I think it would be helpful to make students include their research papers, because we are assessing academic writing (at least in level 4) and that is more realistic than some of the other papers.

3. How effective do you feel the LAT calibration meetings are? Do you have any suggestions for the improvement of these meeting?

I think they're very necessary, and having good benchmarks to look at is essential. I also think it works best when teachers are working together throughout the semester so when we get to portfolios we're already on the same page (like with Writing 4 Club).

4. Do you have any other suggestions for the improvement of the LATs?

Other than improving the scale, which would require us to be familiar with benchmarks at various levels, maybe requiring certain papers in the portfolios would be good, so we can see if we have reached certain objectives.