Second Language Writing Portfolio Assessment

The Influences of the Assessment Criteria and the Rating Process on Holistic Scores

BY CRAIG JAMES CONRAD

CARLA Working Paper #20

CARLA | Center for Advanced Research on Language Acquisition

University of Minnesota

Second Language Writing Portfolio Assessment:

The Influences of the Assessment Criteria and the Rating Process on Holistic Scores

Craig James Conrad

The Center for Advanced Research on Language Acquisition
University of Minnesota, Minneapolis

October, 2001

Second Language Writing Portfolio Assessment

The Influences of the Assessment Criteria and the Rating Process on Holistic Scores

First Edition, Second Printing

© 2001, 2013 by the Board of Regents of the University of Minnesota. All rights reserved.

Produced by

Center for Advanced Research on Language Acquisition
University of Minnesota
140 University International Center
331 17th Ave SE
Minneapolis, MN 55414
USA
612.626.8600
carla@umn.edu
http://www.carla.umn.edu

TABLE OF CONTENTS

LIST OF TABLES iii
ABSTRACT v
1. INTRODUCTION AND REVIEW OF THE LITERATURE 1
1.1 Justification for writing portfolio assessment 1
1.2 Psychometric issues in portfolio assessment 4
1.3 Holistic scoring, the portfolio rating process, and the scoring rubric 6
1.3.1 Holistic scoring in writing assessment 6
1.3.2 The influences of different writing characteristics on holistic scores 7
1.3.3 Portfolios and the rating process 10
1.3.4 The scoring rubric 11
1.4 Motivation for the study and research questions 11
2. RESEARCH DESIGN 13
2.1 Participants 13
2.1.1 Survey participants 13
2.1.2 Verbal report participants 13
2.2 Instrumentation 14
2.2.1 Portfolio assessment program 14
2.2.2 Rater questionnaire 16
2.2.3 Verbal report 17
2.3 Data collection procedures 17
2.3.1 Rater questionnaires 18
2.3.2 Verbal report collection 18
2.4 Data analysis procedures 20
2.4.1 Rater questionnaire analysis 20
2.4.2 Verbal report analysis 20
3. RESULTS 23
3.1 Research question #1: The influence of various criteria on holistic scores 23
3.2 Research question #2: How raters' perceptions of the relative importance of various criteria match the use of those criteria in actual ratings 27
3.3 Research question #3: The effect of the portfolio assessment process on scoring outcomes 28
4. DISCUSSION 31
4.1 Interpretation of results and implications for portfolio assessment 31
4.2 Limitations of the study 33
4.3 Suggestions for future research 35
4.4 Pedagogical implications 36
5. CONCLUSION 39

REFERENCES 41
APPENDIX A: Portfolio Scoring Rubric 44
APPENDIX B: Rater Questionnaire 45
APPENDIX C: Verbal Report Guidelines 50
APPENDIX D: Summary of Pilot Study Design and Results 51

LIST OF TABLES

Table 2.1 Formula for determining overall portfolio scores 16
Tables 2.2-2.4: Assessment criteria categories for the three writing types..... 21
Table 2.2 Portfolio letter criteria categories 21
Table 2.3 Multiple-draft essay criteria categories 21
Table 2.4 Unassisted writing criteria categories..... 21
Tables 3.1-3.3: Frequency of criteria considered for ratings of each writing type..... 23
Table 3.1 Frequency of criteria considered in rating portfolio letters..... 23
Table 3.2 Frequency of criteria considered in rating multi-draft essays..... 23
Table 3.3 Frequency of criteria considered in rating unassisted writings..... 24
Tables 3.4-3.6: Frequency of writing deficiencies by criteria category for *marginal* and *unacceptable* ratings in each writing type 26
Table 3.4 Frequency of writing deficiencies by criteria category for *marginal* and *unacceptable* portfolio letter ratings 26
Table 3.5 Frequency of writing deficiencies by criteria category for *marginal* and *unacceptable* multi-draft essay ratings..... 26
Table 3.6 Frequency of writing deficiencies by criteria category for *marginal* and *unacceptable* unassisted writing ratings 26
Tables 3.7-3.9: Raters' perceptions of relative importance of rating criteria..... 27
Table 3.7 Relative importance of rating criteria for portfolio letters..... 27
Table 3.8 Relative importance of rating criteria for multi-draft essays..... 27
Table 3.9 Relative importance of rating criteria for unassisted writings..... 28
Tables D1-D3: Frequency of criteria considered for ratings of each writing type (Pilot study)..... 51
Table D1. Frequency of criteria considered in rating portfolio letters (Pilot study)..... 51
Table D2. Frequency of criteria considered in rating multi-draft essays (Pilot study)..... 51
Table D3. Frequency of criteria considered in rating unassisted writings (Pilot study)..... 52

Abstract

This paper reports on a study of the influences of the assessment criteria and the rating process on holistic scores assigned to second language writing portfolio components. The study was conducted in the context of an English as a Second Language program for adult international students. Fifteen raters from this program participated in the initial survey phase, in which they were asked to rank the program's various assessment criteria according to their relative importance in determining writing quality. In addition, four of the most experienced raters provided a verbal report while participating in a holistic rating session in which they each rated five different portfolios submitted by ESL students from an advanced-level composition class. The verbal report was examined in order to determine the criteria by which the raters judged each writing sample in the portfolios as well as to identify any effects that the portfolio scoring process had on scoring outcomes. Both the survey and verbal report results indicated that raters were most influenced by the content of the writing samples. The criteria of organization and language usage varied in their relative influence depending on the writing type being assessed. The verbal report also revealed that readers may engage in both bottom-up and top-down rating behavior, and that both of these processes can influence assigned scores. The findings of the study raise questions as to whether holistic scoring is the most valid scoring procedure in assessing the variety of writing found in portfolios.

1. Introduction and Review of the Literature

Portfolio assessment continues to become increasingly widespread as both first and second language writing programs discover its potential as a means of evaluating writing proficiency. However, a number of problematic issues regarding portfolio assessment have gone largely unresolved. Among these issues lie questions concerning the scoring procedures used and the rating behavior that readers exhibit while assessing the quality of writing portfolios. Without a doubt, these are areas which need to be investigated in order for portfolio assessment to be fully justified and validated as a meaningful measure of writing ability. This paper reports on a study investigating both the process and product of assessing second language writing portfolios using a holistic scoring procedure. Prior to the report on this particular study, a preliminary justification for portfolio assessment is presented, relevant issues regarding validity and reliability are addressed, and existing literature on holistic writing assessment and the rating process is reviewed.

1.1 Justification for writing portfolio assessment

The field of educational assessment has evolved over the past few decades as a result of the emergence of current theories of learning and education. Whereas learning was once thought of as a linear progression of acquired knowledge and skills, it is now seen more as a complex, nonlinear process that involves dramatic and intermittent changes in the learner's understanding and ability (Wolf, Bixby, Glenn, & Gardner, 1991). Because of this shift in the way that learning is viewed, it has become a goal of educational assessment to develop instruments that better measure learning in light of its complexity and nonlinear nature.

The use of portfolios as a means of assessing writing in both first and second language contexts has emerged mainly in response to a general dissatisfaction with more traditional forms of writing assessment that were developed during that period in which learning was conceptualized differently. Before the 1970s, the quality of student writing was typically assessed through the use of indirect (and usually multiple-choice) tests of usage and mechanics (Huot, 1994). In accordance with a somewhat antiquated theory of learning, these tests were founded on the underlying assumption that the ability to write is fundamentally governed by the linear acquisition of a discrete set of skills. From an educational measurement standpoint, the presumed advantages of these types of tests included the notion that they could be objectively and reliably scored. However, in more recent times, a number of objections have been raised regarding the

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

questionable validity of such indirect assessments. Most notably, it has been argued that indirect tests of writing lack validity because they do not accurately represent the construct of writing, or in other words, what it means to write.

The negative reactions to the traditional indirect writing tests led to the use of more direct assessments in the form of timed-essay exams, which were considered to be more realistic reflections of the construct of writing. However, an initial problem discovered with the essay exams was that they could not be scored as reliably as the indirect writing tests. In addition, a number of other objections have called into question the use of the timed-essay assessment. Some of these issues are related to potential problems in the test design, such as prompt development and time constraints (Hamp-Lyons & Kroll, 1996). Other issues have been raised regarding the validity of these tests. Perhaps the most basic argument against the validity of timed-essay exams is that they measure abilities such as "a quick memory, fluency, [and the] ability to turn out reasonably clean and organized first draft work to someone else's topic under time pressure" (White, 1994, p. 33) instead of the true ability to write. Another claim against the validity of timed essays is that in eliciting only a single writing sample, they do not sufficiently measure the abilities that students must demonstrate in order to succeed on the various writing tasks found throughout the different academic disciplines (see Horowitz, 1991, for a discussion). Related to this is the argument that timed exams of writing per se do not truly reflect the purposes of the essay tests that are used in academic content classes (Armstrong Smith, 1991), which is significant in light of the fact that the mere existence of essay exams in academic courses has been seen as a justification for their use in writing assessment. Furthermore, some research suggests that second-language writers are particularly disadvantaged when it comes to timed-essay exams. For example, in examining the pass rates of both ESL and native English-speaking students on an institutional exit proficiency exam, Ruetten (1994) found that although ESL students were twice as likely to fail the exam, evidence of their writing ability demonstrated through an appeals process ultimately led to a pass rate that was comparable to the native-speaking students. Similarly, evidence from another study pointed out a significantly high failure rate on timed-essay exams for nonnative students who were otherwise academically successful (Byrd & Nelson, 1995).

The use of writing portfolios as assessment instruments has been hailed to a certain extent as a potential answer to the shortcomings of both the indirect writing test and the more direct timed-essay assessment. Portfolios share the common goal of other "alternative, authentic, or performance" assessments, which is essentially to provide evidence regarding the complex processes in which students engage themselves in actual, real-life performances (Camp, 1993; Brown &

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

Hudson, 1998; Gitomer, 1993; Huerta-Macias, 1995; and Linn, Baker, & Dunbar, 1991). While an all-encompassing definition of the writing portfolio is difficult to arrive at, the portfolio programs used at many institutions seem to share a number of commonalities which will be used to operationalize the term *portfolio* for the remainder of this paper. They include the following:

1. Multiple samples of writing gathered over a number of occasions.
2. Variety in the kinds of writing or purposes for writing that are represented.
3. Evidence of process in the creation of one or more pieces of writing.
4. Reflection on individual pieces of writing and/or on changes observable over time.

(Camp & Levine, 1991, p. 197)

Although not the case for all forms of portfolio assessment, a fifth characteristic to be included in this definition due to its fundamental importance to the current study is that a portfolio is read and assessed by more than one rater. Finally, since the process of compiling and assessing portfolios involves taking samples of work from the classroom itself and having this work evaluated by instructors from the institution in which it is produced, portfolios might be referred to as a type of "contextualized performance assessment" (Camp, 1993, p. 186).

The potential benefits of the use of writing portfolios can be described in terms of the ability to enhance not only the process of writing assessment, but also student learning and the role of the teacher (see Brown & Hudson, 1998 for a detailed discussion). According to White (1994), "portfolios bring teaching, learning, and assessment together as mutually supportive activities, as opposed to the artificiality of conventional tests" (p. 27). Since the current paper's focus is on portfolios as an assessment tool, the advantages they offer in this regard are of particular interest. Several published articles have documented the positive effects on the assessment process that portfolios have had when introduced into the first language writing programs at a number of institutions. For example, the implementation of portfolio assessment at the University of Michigan is reported to have significantly contributed to a growing sense of program consensus in defining the construct of writing competence as well as in shaping the program's instructional curriculum (Condon & Hamp-Lyons, 1991). At another large midwestern university, an increased teacher influence on the assessment process, the promotion of high standards and scoring consistency among teachers, and an increased internalization of writing assessment standards by beginning instructors have been reported as beneficial results of a newly adopted portfolio assessment system (Roemer, Shultz, & Durst, 1991). Finally, at SUNY-Stony Brook, portfolio assessment has been found to better recognize the intricacies involved in the various stages of process writing (Elbow & Belanoff, 1986).

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

Many other advantages of portfolio assessment in general are referred to in the literature, but what are perhaps of most significance for the purposes of this paper are the advantages of portfolio assessment for second language writers in particular. Citing evidence from her experience at the University of Michigan, Hamp-Lyons (1996b) discusses how the introduction of portfolios, along with the elimination of the timed-writing assessment context, has allowed more ESL students to test out of the lowest-level mainstream writing class upon their first attempt. She believes that one of the reasons behind this is the increased amount of time for revisions, which allows nonnative students the opportunity to correct any "fossilized errors" that might have otherwise surfaced and gone unrevised under the unnatural time constraints of an essay test. It is also claimed that nonnative writers benefit, in a portfolio-based assessment context, from having the chance to revise the various aspects of their writing (e.g., idea development, organization, grammar/mechanics) at different stages in the drafting process, instead of having to attend to competing textual needs at the same time (Hamp-Lyons & Condon, 2000). Another advantage of portfolio assessment for NNS writers is its multi-dimensional nature. Timed-essay tests, which elicit only one relatively small writing sample, may not reveal the complete picture of the writer's abilities. Portfolios, on the other hand, with their greater number of texts and multiple drafts, may provide a more comprehensive and consequently fairer assessment of a nonnative writer's ability (Hamp-Lyons, 1995a).

1.2 Psychometric issues in portfolio assessment

In spite of the benefits of portfolio assessment, a number of problematic issues have emerged regarding the use of portfolios in measuring writing ability. These areas of contention are usually described in terms of the questionable psychometric properties (i.e., validity and reliability) of portfolios as an assessment tool. There are those who insist that even though alternative assessments may not mesh well with traditional psychometric criteria, these criteria must still be satisfied, especially when the results of the assessments are used for high-stakes purposes (Miller & Legg, 1993). Others argue in favor of the need to reconceptualize and/or expand such notions as validity and reliability in light of the complex nature and goals of such alternative assessments as portfolios (Wolf et al., 1991; Linn et al., 1991; Camp, 1993). The issues surrounding validity and reliability are of great importance to the present study since its purpose is to examine the rating process, which has direct impact on the validity and reliability of the assessment.

The notion of validity, commonly defined as the extent to which an assessment measures what it intends to measure, is often spoken of in terms of its many different aspects. In order to

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

fully validate performance assessments such as portfolios, Linn et al. (1991) suggest eight criteria that must be satisfied: consequences, fairness, transfer and generalizability, cognitive complexity, content quality, content coverage, meaningfulness, and cost and efficiency. Unfortunately, as Hamp-Lyons (1996a) points out, research into portfolio assessment is still very limited in all eight of these categories. Of all the evidence available regarding the different types of validity, the strongest argument in favor of portfolios is that they exhibit a high degree of face validity—meaning their appearance reflects what they claim to assess—since they reflect the natural writing process and consist of writing that is produced in authentic contexts (i.e., non-assessment contexts). However, face validity is commonly thought of as the least important aspect of validity mainly because it is not rigorously determined by using theoretically or empirically established criteria (Anastasi, 1976; Lyman, 1978; Popham, 1981; and White, 1985—all cited in Huot, 1990). Therefore, serious investigations are still needed if portfolios can be fully considered to be valid assessment tools.

In addition to validity, the achievement of acceptable levels of reliability has traditionally been thought of as a necessary condition for any meaningful assessment, with the reasoning being that an assessment is only as valid as it can be reliably scored and interpreted. Specifically, this argument claims that if the results of a particular assessment are found to be unreliable, they are basically meaningless since they can not be said to be indicative of future assessments or performances in the skill being tested (Huot, 1990). However, there is no consensus of opinions on this issue in the area of portfolio assessment, especially as it is concerned with interrater reliability in particular. For example, on one extreme there are proponents of portfolio assessment who believe that lack of agreement among raters, although detrimental to reliability, is beneficial to the assessment process since it can provide useful information about the texts being assessed and the readers conducting the assessment (Broad, 1994). Along this same vein, Shale (1996), in referring to writing assessment in general, proposes that we overlook interrater reliability to some extent and use current generalizability theory to account for the natural variance in raters' judgment of ability. Elbow (1991) goes so far as to claim that validity and reliability are actually in conflict, and that achieving a valid assessment justifies a certain level of neglect of reliability.

Although these arguments in favor of downplaying the importance of reliability in portfolio assessment may have their merits, reliability can not be ignored when determining the value of this form of assessment. For the sake of fairness, one of the conditions of validity established by Linn et al. (1991), an acceptable degree of reliability is particularly important in situations where the outcomes of the assessment are used to make high-stakes decisions, such as in the case of an exit

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

assessment. Therefore, it is important to ensure that raters are consistent in agreeing on the quality of writing in a given portfolio. Admittedly, this is not an easy chore in light of the number of complex factors that influence a rater's judgment, including the characteristics of the assessment task, characteristics of language, the actual performance on the task, the criteria established for the assessment, and the rater's own personal criteria (Reed & Cohen, 2001). The fact that the assessment task itself varies within and across portfolios as a result of the unique nature of each portfolio makes the job of ensuring interrater reliability even more difficult than in other forms of assessment.

1.3 Holistic scoring, the portfolio rating process, and the scoring rubric

The focus of the present study's investigation is of fundamental importance to the process of determining the validity and reliability of portfolio assessment for second language writers. In addition to the task involved and the writer himself, both the scoring procedure and the reader(s) of the writing samples must be taken into account when validating any form of writing assessment (Hamp-Lyons, 1990). This study, in examining the criteria that raters use in holistic assessments of portfolios, is particularly concerned with the scoring procedure as well as the rating behavior of readers and the ways in which this behavior is influenced by the various textual features found within the multiple writing samples contained in a portfolio. Therefore, it is necessary to look at the issues that have been raised regarding holistic scoring, the findings of previous research into holistic scoring, the process of assessing writing portfolios in particular, and the scoring rubrics that are used to guide the readers in their assessments.

1.3.1 Holistic scoring in writing assessment

Four general types of scoring methods are available for use in assessing writing: holistic, analytic, primary trait, and multiple trait (for a description of these, see Cohen, 1994). Of these four types, holistic scoring has emerged as perhaps the most common as a result of its relatively high interrater reliability coefficients (Huot, 1990). This scoring method can most basically be defined as one which assigns "a single grade based on the total impression of a composition as a whole text or discourse" (Perkins, 1983, p. 652). In addition, holistic scoring often involves two readers and a possible third in situations where the initial readers disagree to an unacceptable extent (Hamp-Lyons, 1995b). In these cases where multiple readers are involved, their scores are often averaged to arrive at a final score. Finally, an important characteristic of holistic scoring is that

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

training in the procedure is provided to the various raters so that they share a sense of the criteria being used and can then apply these criteria consistently.

The rationale behind adopting a holistic scoring approach to writing assessment is that overall writing quality can not be determined by merely counting point values assigned to various elements of a given text, and thus it is necessary to judge a sample of writing as a whole instead of in terms of the sum of its parts. Because of this fundamental characteristic of holistic scoring, it has been claimed to possess the greatest degree of construct validity of all the scoring procedures in measurement of general writing proficiency (Perkins, 1983). However, this assertion is not accepted universally, and a number of potential disadvantages to holistic scoring have also been cited. With reference to nonnative writers in particular, holistic scores are not informative about differing levels of quality on different writing traits (e.g., content and grammar), thus ignoring potentially useful information about the writer's ability in these different areas (Hamp-Lyons, 1995b). In other words, holistic scoring procedures are ultimately non-communicative, in the sense that they do not say much about the writer's specific strengths and weaknesses. Of all the arguments against holistic scoring, perhaps the most important question has to do with its validity as a method of assessment. It is not clear whether rater training in fact ensures that readers will consistently apply the same criteria in arriving at holistic judgments of writing quality. Furthermore, while the criteria prescribed for use by the raters are obviously accepted by those who develop them, they are not necessarily accepted by those who are expected to use them in actual reading sessions (Charney, 1984). In such cases, a lack of acceptance of criteria by the raters seriously jeopardizes the validity of the assessment since such raters are likely to be idiosyncratic in their rating behavior.

1.3.2 The influences of different writing characteristics on holistic scores

Because the validity of holistic scoring is dependent on the consistent application of rating criteria, it is important to determine exactly what criteria raters use in arriving at their final scores. In light of the fact that holistic scoring typically condenses a number of criteria (e.g., content, organization, grammar/mechanics) into composite descriptions of proficiency, some research has uncovered evidence regarding the relative influence of different textual features in arriving at overall holistic scores. At least two of these studies have examined and compared holistic and analytic scores assigned by both English (native-language) and ESL instructors to the writing of both native and nonnative speakers of English. In one of these studies (O'Loughlin, 1994), four experienced raters from each of these two groups read and scored 20 native-speaker and 20

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

nonnative-speaker essays, using both holistic and analytic scoring procedures. The results indicated that holistic scores for both native-speaker and ESL essays were most influenced by the two analytic categories of *arguments & evidence* and *organization*. This finding held true for the raters from both groups. There was not substantial evidence to describe the influence of the other three analytic categories of *appropriateness*, *grammar & cohesion*, and *spelling & punctuation*.

In another study, Song and Caruso (1996) examined the degree of difference in ratings performed by 32 English and 30 ESL professors on two essays written by native speakers of English and two written by ESL students. Half of the members of each group of raters used a holistic scale and the other half of the members of each group used an analytic scale comprised of 10 different textual features, six of which were identified as rhetorical features (e.g., clearly stated or reasonably implied central focus, supportive elements that clarify the central focus, use of traditional and/or organic transitional elements) and four of which were identified as language usage features (e.g., overall control of language, variety and complexity in sentence structure). With respect to the relative influences of different textual features on holistic scores, the authors concluded that the English teachers gave most weight to what they referred to as overall content and rhetorical features, as opposed to language usage, for both native-English and ESL writers. Unfortunately, no such findings were reported with regard to the influence of different features on the ratings given by the ESL teachers.

A recent study by Chiang (1999) provides evidence that readers of second language essays are more likely to be influenced by what he refers to as discourse features (i.e., *cohesion* and *coherence*), as opposed to language usage (i.e., *morphology* and *syntax*), when assigning holistic scores. In this study, three native speakers of French rated 172 essays from students of French as a foreign language on an analytic scale covering four areas of evaluation (*morphology*, *syntax*, *cohesion*, and *coherence*) in addition to assigning a single holistic score. The four analytic categories included 35 different features to be assessed individually. The results indicated that *cohesion* (e.g., appropriate and accurate use of pronouns of reference, appropriate use of ellipsis, judicious and accurate use of junction words) was the most influential feature in assessing overall quality, followed by *coherence* (e.g., relevance of ideas to the topic, relationship of ideas to one another, elaboration of ideas, smoothness of transitions between paragraphs). However, it is interesting to note that although cohesion was the most influential feature, it was also the feature with which raters disagreed most in terms of quality.

The results of at least one study contradict the findings that raters are most influenced by content and rhetorical features of L2 writing. Research conducted by Sweedler-Brown (1993)

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

examined the holistic ratings of ESL essays in particular as given by English composition teachers who had not been trained in ESL. Six ESL essays from intermediate-level university students were chosen, and they were subsequently rewritten with the typically nonnative sentence-level errors corrected. Each of six raters scored three original and three rewritten essays holistically and also on a separate analytic scale, with both scales ranging from 1 to 6. Unlike the results of the previously discussed studies, Sweedler-Brown found the language usage features of *sentence structure* and *grammar/mechanics* to be more influential than the rhetorical features of *organization* and *paragraph development* in determining overall holistic scores. In fact, no correlation was found between the analytic scores on the latter two features and the holistic scores assigned to the original and rewritten essays. The researcher explains this finding as a possible result of lack of training in ESL on the part of the raters.

What is perhaps the greatest limitation of the previously described studies is the fact that they assume a rather direct relationship between holistic and analytic scoring procedures. In other words, these studies have assumed that it is plausible to make correlations between individual scores on various analytic criteria and a single holistic score that is not arrived at through any formula of combining the analytic scores. However, it can be argued that holistic and analytic methods are fundamentally different and that any correlations made between them must be done so with some acknowledgement of limitations. For example, while reflecting on the construct validity of his analytic scale, Chiang (1999) mentions the likely possibility that the scale itself, with all its discretely defined categories and features, may have actually forced raters to assess aspects of writing that they would not do under other circumstances, as in a strictly holistic assessment situation. Therefore, studies utilizing the type of methodology which correlates holistic and analytic scores must be complemented with others that involve research procedures that attempt to simulate authentic holistic assessments.

One such study by Vaughan (1991) involved think-aloud protocol analyses in order to better understand the thought processes raters engage in during holistic writing assessments. Nine readers experienced in holistic assessment rated six essays on a six-point holistic scale. Two of the essays were written by native speakers of English and four were written by nonnative speakers. The raters were asked to tape-record their comments as they read, with specific instructions to act as they would in normal holistic rating sessions. Rater comments were transcribed in detail, and each meaningful unit was labeled and sorted into 14 general categories. Analysis of the data revealed that of the 14 different categories, *content* was the category most frequently commented on by five of the nine raters, followed by *grammar* (3 raters, including one who commented equally on grammar

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

and content) and *organization* (2 raters). While the majority of rater comments fell into these three general categories, Vaughan noted that some specific and salient characteristics of certain essays made strong impressions on the raters. For example, handwriting was one of the most commonly mentioned shortcomings of the essays. This finding is significant since some of these characteristics, including handwriting, were not accounted for in the scoring guidelines, and therefore may reflect personal approaches to holistic assessment. Moreover, as a result of a high degree of this type of idiosyncratic behavior among raters, many would argue that the validity and reliability of the assessment instrument can be negatively affected.

1.3.3 Portfolios and the rating process

While the studies of holistic writing assessment cited above are concerned with single-sample essays, very little has been done to address the rating process involved in the assessment of writing portfolios, regardless of the scoring method used. Perhaps the most fruitful work that has emerged in this area is that of Hamp-Lyons and Condon (1993). Through their work with portfolio readers at the University of Michigan, they sought information regarding, among other things, how and when judgments were reached on the quality of a given portfolio and the standards that the readers used in reaching these judgments. However, instead of arriving at concrete answers to these areas of inquiry, they were led to question a number of assumptions they had previously held regarding writing portfolio assessment.

Based on reader surveys, one conclusion they reached which is somewhat relevant to this paper is that it is quite doubtful that a portfolio can be assessed holistically as a single entity. Instead it is much more likely that readers will assess the component texts individually and weigh each of them in light of the others in order to come to a final decision on the portfolio's quality. It was also quite clear that readers oftentimes do not consider all texts and components of the portfolio equally, and more alarmingly, readers may arrive at a judgment of a given portfolio's quality without having read all of its components. Another concern, which was raised by readers and said to influence their assessments, was that some improvements found among multiple drafts of a text were likely the result of the instructor rather than the writer. All of these issues are of obvious importance in examining the process of assessing writing portfolios since they can affect the scoring outcomes. Finally, the authors underscore the importance of regularly conducting standardization sessions, in which the instructors can collaborate in order to identify and define the criteria that should be used to assess the portfolios.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

1.3.4 The scoring rubric

In addition to the scoring procedure, be it holistic or otherwise, there are issues related to the design of the scoring rubric itself which affect the rating outcomes and must be accounted for when validating a particular form of writing assessment. Obviously, designing a rubric for a single-sample writing assessment on a given topic is much less complicated than designing a rubric that has the purpose of describing levels of writing ability on the various genres of writing found in portfolios, not to mention the lack of standardized topics within them. In some cases, such as when only a few genres of writing are represented in the portfolios, a somewhat specific rubric can be used. In other situations where a greater number of genres appear, the descriptors on the rubric must be general enough to reflect the variety of writing that is being assessed. A significant problem with a general rubric is that the language found in the descriptors, such as a term like *clarity*, is inevitably difficult to define (Callahan, 1995). Furthermore, it is likely to be equally challenging to draw clear distinctions between the standards of quality to which the descriptors refer. For example, it is quite difficult to delineate the boundary between *very clear organization* and *somewhat clear organization*. On the other hand, more specific scoring rubrics are not without problems of their own. While they may lead to greater interrater reliability because they are usually more easily followed, they may end up attending to writing qualities that are superficial (Wiggins, 1994, in Callahan, 1995), which is likely to damage the assessment's validity.

1.4 Motivation for the study and research questions

While the available research on holistic writing assessment has provided some evidence about factors that affect scoring outcomes on single-sample compositions, the general lack of information regarding the assessment of portfolios and the criteria raters use in judging their quality has become the primary motivation for the current study. According to Hamp-Lyons and Condon (2000), the combination of shared criteria and standards among raters is one of seven crucial characteristics of any successful portfolio-based assessment. In any type of writing assessment, it is critically important to examine the criteria used by the various raters in order to determine the construct validity of the assessment (Hamp-Lyons, 1990). If raters exhibit tendencies to apply the same criteria in their assessments, then it can be assumed that they share the same construct of writing quality. If they are overly idiosyncratic in the criteria they use, then a single construct of writing quality is not shared among the group, and the assessment loses its meaning and validity. Furthermore, whenever a holistic scoring procedure is adopted, it is important to determine whether raters indeed arrive at scores through holistic judgments of overall

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

quality, or if instead their assessments are overly influenced by certain, specific aspects of the writing. While certain criteria are undoubtedly more important than others, the theory behind holistic assessment is based on the assumption that writing should be assessed according to its merits as a whole text. Therefore, it is in the best interests of any holistic assessment to provide evidence that all of the rating criteria are considered to a reasonable extent. In order to investigate these issues and begin to determine the validity of holistic scoring in portfolio assessment, it is necessary to determine the relative influence of various criteria on the holistic scores that are assigned.

Obviously, the reading/rating process has a very direct impact on the product of the assessment, or the scores that are given to the writing samples. Therefore, it is also extremely important to examine this rating process in order to understand the ways in which readers arrive at judgments of writing quality. Otherwise, if the rating process is left unaccounted for, it is impossible to validate the judgments that are rendered on the writing samples. In other words, "if we do not know what raters are doing (and why they are doing it), then we do not know what their ratings mean" (Connor-Linton, 1995, p. 763). Without question, the scoring procedure used (e.g., holistic scoring) as well as the content and format of the scoring rubric will affect the ways in which raters go about their task and will ultimately influence the outcomes of their assessments.

In order to better understand the assessment of second language writing portfolios, this study asks the following research questions:

1. To what extent do various scoring criteria influence raters' holistic assessments of second language writing portfolio components?
2. How well does the actual use of these criteria in rating match the raters' perceptions of their relative importance?
3. What effect does the unique nature of the portfolio assessment process have on scoring outcomes?

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

2. Research Design

The overall design of this study consisted of two separate stages: 1) a survey stage, in which a questionnaire was completed by fifteen raters from an intensive English as a Second Language program for international students; and 2) a verbal report stage, in which four of these fifteen initial participants provided think-aloud commentary while reading and assessing the contents of second language writing portfolios. Furthermore, this verbal report stage included two different data collection sessions—a pilot study and the main study. The following sections of this report describe the design of the different phases of the study in further detail.

2.1 Participants

2.1.1 Survey participants

For the initial survey phase of the study, fifteen individuals with experience in the ESL writing portfolio assessment program described below agreed to participate. Eleven of these participants were involved in the portfolio assessment program at the time of the study. The other four participants had been involved in the same portfolio program within the three years prior to the study but were no longer involved at the time the research was conducted. In addition to regularly serving as portfolio readers, two of the participants were also the administrators of the portfolio assessment program. The ages of the participants ranged from the mid-twenties to over fifty years, and all but two of the individuals were female. At the time the survey was conducted, fourteen of the fifteen participants reported having six or more years of experience teaching ESL, and ten reported at least six years of experience teaching composition to ESL students. With respect to experience in portfolio assessment, eleven of the fifteen participants had taken part in six or more of the nineteen portfolio reading sessions in the history of the program to date.

2.1.2 Verbal report participants

For the second phase of the study, four of the most experienced raters involved in the portfolio assessment program at the time of the research participated. These four raters had also participated in the rater survey approximately two months prior to their involvement in this stage of the study. One of these individuals was serving as both a portfolio program administrator and a portfolio reader/rater at the time of the study. All of the four participants had taught ESL composition for at least six years, and all of them had taken part in at least sixteen of the nineteen

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

portfolio assessment sessions in the program's history. Three of the four participants were female, and the ages of the four individuals ranged from thirty to over fifty years.

2.2 Instrumentation

2.2.1 Portfolio assessment program

This study was conducted in the context of an intensive English as a Second Language program for international students at a large midwestern university. In this program, portfolios were used as one indicator of writing proficiency for advanced-level students. At the end of each term, these students submitted portfolios containing three different writing samples. The first of these items was an introductory letter addressed to the portfolio readers, which not only served the function of introducing the portfolio to the reader, but was also an assessed component of the portfolio. This cover letter was described to the students as having the following four purposes (reproduced verbatim from a program handout):

1. To explain the contents of your portfolio to readers who do not know anything about you or your class.
2. To explain the assignments for the work in your portfolio.
3. To show self-awareness as a writer—that is, to show you have the ability to reflect on your writing.
4. To show that you can write a clear, well-organized letter. (The letter itself will be evaluated, just like the other papers in the portfolio.)

The other two portfolio writing components were a multiple-draft essay and an unassisted writing, both of which the students were free to select at their own discretion. The multiple-draft essay was typically the most developed writing sample since it was a piece that had gone through various iterations and had benefited from teacher feedback throughout its various stages. In addition to the final version of the multi-draft essay, the students were required to submit all prior drafts in the portfolio, which typically would result in a total of three to four drafts including the final version. Along with the different drafts, the students were also required to include any comments that were provided by the instructor regarding the essay in its various stages. The unassisted writing was essentially a piece that the student had written without the assistance of the instructor. This item was typically a rewritten journal entry or a commentary about a reading, although in the latter case, students were expected to go beyond simply summarizing the reading.

Following the completion of each term, a group of instructors would meet to read and assess these portfolios as a testing responsibility required by the program. Before the actual reading

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

session, the readers typically participated in a standardization meeting in which they discussed the quality of a set of anchor portfolio writings chosen by the portfolio program administrators. This discussion was intended to build consensus on the standards that were to be used to distinguish portfolios of differing quality based on the descriptors found on the portfolio scoring rubric. In order to set standards, the group would discuss both previously submitted portfolio writings as well as newly submitted items. Standards were then set democratically according to the majority of the group's opinion of each sample's quality. After this discussion, each submitted portfolio was read and rated by two instructors, and in cases where there was disagreement between them, a third reader was included in the assessment. Readers who taught an advanced composition class during the term were not to assess portfolios from students in their own class.

The scoring rubric which was used to assess the portfolio items was composed of a separate set of descriptors for each of the three types of writing (see Appendix A). In other words, the portfolio letter, multi-draft essay, and unassisted writing were all assessed according to somewhat different criteria. The sections of the rubric used to assess both the portfolio letter and the unassisted writing involved holistic procedures in the traditional sense of using composite descriptions of overall quality. The section of the rubric intended for assessing the multi-draft essay was not purely holistic in the sense that the descriptors for the different traits (i.e., content, audience awareness, organization, and language usage) were described separately. In this way it resembled a multiple-trait scoring instrument to a certain extent (see Hamp-Lyons, 1991). However, the procedure for scoring the multiple-draft essays, like the other two components, was essentially holistic since the reader assigned a single overall score for the essay based on an impression of its overall quality, without any specific formula for arriving at that score.

All three writing samples were scored on a three-level scale: *acceptable*, *marginal*, or *unacceptable*. The overall portfolio score, also reported according to this same scale, was then arrived at by means of a given formula (see Table 2.1). According to this method of determining the final portfolio grade, if any one writing sample received a score of *unacceptable*, then the overall portfolio grade was *unacceptable*. Either of the following two situations would result in an overall portfolio grade of *marginal* at best: 1) if the multi-draft essay was scored as *marginal*; or 2) if the portfolio letter and unassisted writing were scored as *marginal*. In cases where either the portfolio letter or unassisted writing was scored as *marginal* and the other two writings were found to be *acceptable*, the overall grade was *acceptable*. Obviously, if all three items were found to be acceptable, that same grade was given to the portfolio as a whole.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

Table 2.1 Formula for determining overall portfolio scores

Individual writing sample scores	Overall portfolio score
One *unacceptable* score	*Unacceptable* portfolio
A *marginal* multi-draft essay	*Marginal* portfolio at best
Two *marginal* scores	*Marginal* portfolio at best
One *marginal* letter or unassisted writing & two *acceptable* scores	*Acceptable* portfolio

Ideally, the final judgment on the quality of a portfolio was arrived at once two readers had assessed it. In situations where the final overall scores arrived at by two readers were equal, then the portfolio received that grade. If the two readers differed in the final scores they assigned, a third reader was called upon to assess the portfolio or certain items within it. If the third reader agreed with one of the previous readers on the overall portfolio grade, then that was the final score assigned to it. In cases where all three readers disagreed, either a fourth reader was called upon to assess the portfolio or some kind of compromise was reached.

As mentioned earlier, the portfolio grade was one measure of the student's writing proficiency at the end of a given term. In determining whether or not a student was ready for the next level of writing instruction or exempt altogether from future ESL composition classes, the results of the portfolio assessment were considered in conjunction with a final recommendation made by the student's composition instructor and scores from the standardized Test of English as a Foreign Language (TOEFL). While the portfolio grade was not the primary influence in making these decisions, a passing portfolio could settle cases in which the other indicators of the student's writing proficiency were borderline. Therefore, the quality of a student's portfolio could be a determining factor in a multifaceted exit assessment of the student's writing ability.

2.2.2 Rater questionnaire

The purpose of the initial survey phase of the research was to obtain information concerning all of the current portfolio readers as well as any instructors who had assessed portfolios in the program within the past three years. A questionnaire (Appendix B) was designed and distributed to the fifteen participants described in Section 2.1.1. The main goal of this survey was to collect data that would describe how the different respondents viewed the relative importance of the various criteria that were intended for use in assessing the three different components of the writing portfolios in the program. In order to obtain this information, the participants were asked to provide a series of numerical rankings indicating how important they felt the various aspects of the

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

writings in the portfolios were in determining overall writing quality. For example, one item asked them to rank the four general criteria used to assess the multi-draft essays (*content*, *audience awareness*, *organization*, *language usage*) according to their opinion of each criterion's importance in determining the essay's overall quality. For this type of item, if a respondent felt that particular factors were of equal importance, he or she was instructed to give equal numerical values to those aspects and adjust the other rankings accordingly. It was felt that this would eliminate the possibility that respondents would be forced to differentiate the importance of criteria which they felt were equally important. Nonetheless, it is worth acknowledging that the mere fact that respondents were asked to rank the various criteria could have biased the results of their rankings to some degree.

2.2.3 Verbal report

The second and more substantive stage of this research involved the collection of verbal report data from the four participants described in Section 3.1.2. As has been done in other similar studies (Huot, 1993; Vaughan, 1991), think-aloud protocol analysis was used to gain insight into the participants' thoughts and judgments during a rating session. The rationale behind using this method of data collection was its relatively low degree of intrusiveness as well as its ability to provide perhaps the most informative evidence about the rating process (Connor-Linton, 1995). In the pilot verbal report stage, the participants read two portfolios, each of which consisted of the three portfolio components described earlier (i.e., portfolio letter, multi-draft essay, and unassisted writing). For the verbal report in the main study, the participants read a series of five portfolios matching the same description. Think-aloud comments were provided by the participants and recorded while they read each text. In addition, upon arriving at a score for each writing sample, the participants were asked to explain all of the reasons for which they assigned that particular score. The comments made in these explanations were to be considered as representative of the criteria the raters used in judging the quality of the writing. Furthermore, in the main study, the raters were asked to identify any single factor for each writing sample that was most influential in affecting their assessment.

2.3 Data collection procedures

2.3.1 Rater questionnaires

The participants were free to complete the questionnaires on their own. The original intention was to give the participants one week to complete the questionnaire. However, because

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

only a few participants responded within this amount of time, the participants were allowed a period of several weeks to complete the questionnaire in order to ensure that as many questionnaires as possible would be returned.

2.3.2 Verbal report collection

All four of the verbal report participants took part in the pilot study, which was conducted approximately two and a half weeks prior to the data collection for the main study. In this pilot, the participants provided think-aloud verbal reports while reading and assessing the components of two portfolios. A summary of the basic design and results of the pilot study is included in Appendix D. In addition to allowing the researcher the opportunity to refine the procedures for the main study, the pilot served the purpose of giving each of the participants an opportunity to gain experience in giving verbal report. It was felt that this would better prepare them for the procedures of the main study and would subsequently lead to a richer body of data.

Four days prior to the data collection for the main study, the four participants took part in a standardization meeting and an actual-stakes portfolio assessment session with the other raters in the program. In the standardization meeting, the quality of four previously submitted portfolio writing samples was discussed in addition to eight newly submitted writing samples. After agreeing on standards, the group read and assessed a set of portfolios which had been recently submitted by students in the program. It was felt that it would be beneficial to collect the verbal report data for the main study after such a standardization meeting since the practice of setting standards is considered to be a crucial component of the assessment.

During the data collection session for the main study, the four participants read and holistically rated the contents of five writing portfolios that had been submitted for a grade at the end of the recently completed academic term. The items in the portfolios were assessed by the participants using the scoring rubric briefly described in Section 2.2.1 and included in its entirety in Appendix A. Each rater was given a photocopied set of the original portfolios and a set of guidelines to follow for giving their verbal report (see Appendix C). The participants were instructed to read and assess the portfolios in the order in which they were arranged on their desks. They were also asked to read and assess the portfolios in a manner as similar as possible to that in an authentic holistic portfolio assessment session. In order to most accurately replicate an actual reading session, the researcher did not intervene during the data collection, nor were the participants given explicit instructions as to what to comment on in order to avoid any bias that could have resulted from influencing raters to employ criteria that they might otherwise disregard. Participants were simply

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

told to verbalize as many of their thoughts as possible while reading, to explain all of the reasons for each score they assign, and to indicate any single factor that was most influential in arriving at a particular score.

All of the portfolios assessed by the raters were compiled by ESL students in the same advanced-level composition course, and no rater had had any of those particular students in his/her own composition class during the term. Therefore, all of the raters entered the rating session without having been previously exposed to these particular portfolios. The five portfolios selected for the main study were chosen based on the composition instructor's opinion of the quality of the writing submitted. These particular portfolios were all judged by the instructor to be of *marginal* (as opposed to *acceptable* or *unacceptable*) quality, according to her interpretation of the portfolio assessment standards. It was felt that these portfolios, being cases of borderline writing, would provide the greatest possibility for a rich array of reader comments and judgments. Each of the portfolios included all three writing types previously described: a portfolio letter, a multiple-draft essay, and an unassisted writing. In accordance with the program's guidelines, all drafts of the essay component were included. All of the texts in the portfolio, with the exception of some of the preliminary drafts, had been typed on a word processor by the students, and any teacher feedback provided on the writing samples or on separate commentary sheets were included in the portfolios, as was normally done in authentic reading sessions. The components of each portfolio were ordered identically, in such a way that the each participant would read the portfolio letter, multiple-draft essay, and unassisted writing in that order.

The rating session was conducted in an audio language lab in which the participants sat in individual booths and wore headsets and microphones for recording purposes. The booths and headsets provided the necessary privacy in order for the raters to participate with the least amount of distraction from other readers as possible. Each participant was seated in a booth in one of the four corners of the room in order to further alleviate any distractions from other participants. The researcher was present in order to ensure that the intended procedures were followed and to make sure that recording equipment was functioning correctly while participants provided the verbal reports. A three-hour period was allotted for the reading session. Two of the participants took approximately one and a half hours to complete their assessments, and the other two participants finished in just under two hours.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

2.4 Data analysis procedures

2.4.1 Rater questionnaire analysis

For the numerical rankings that the fifteen survey participants were asked to complete, descriptive statistical analyses were performed. For each individual item ranked, the mean, median, and standard deviation were calculated. These statistics were used to collectively determine the respondents' perceptions of the relative importance of the various scoring criteria in determining overall writing quality.

2.4.2 Verbal report analysis

The audio tapes of the four raters' verbal reports were transcribed in full with attention paid to emphasis, pause-fillers, and emotional reactions (e.g., laughter). The transcripts were then examined in order to determine the criteria which the raters focused on when considering or explaining a final score for each individual rating. The criteria mentioned as the raters considered the overall quality of each writing sample were noted. These notations did not include specific comments made while reading, but rather only those comments that were considered to be representative of the rater's overall impression of the writing sample. This distinction was made since it was felt that specific comments made about certain local features of the writing may or may not have been considered when deciding the overall final score.

Tables 2.2-2.4 illustrate the various categories of criteria to which the raters' comments could pertain, according to the scoring rubric used in the assessments. These categories were derived from the descriptors on the scoring rubric included in Appendix A. Decisions regarding how to categorize rater comments according to these criteria distinctions were based mainly on how the raters themselves interpreted the criteria and the researcher's understanding of the various criteria. For example, if a rater referred specifically to one of these categories when making a observation and that identification matched the researcher's understanding of the criteria, such a comment was categorized accordingly. In cases where a rater did not identify a comment in terms of one of the criteria categories or there was uncertainty in whether the comment in fact matched the criteria identified by the rater, the researcher consulted the context of the verbal report transcript, the scoring rubric, and the writing sample to which the comment pertained in order to determine into which of the categories the comment best fit. For a handful of comments that were found to be particularly problematic, one of the portfolio assessment administrators was consulted in order to categorize those comments more reliably.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

Tables 2.2-2.4: Assessment criteria categories for the three writing types

Table 2.2 Portfolio letter criteria categories

1. Explanation of assignments and contents of portfolio
2. Awareness of self as a writer (reflection) or clarity of voice
3. Length
4. Organization
5. Language usage (vocabulary and grammar)

Table 2.3 Multiple-draft essay criteria categories

1. Content 1a. Clarity, development, and support of ideas 1b. Clarity of focus 1c. Unity of the essay 1d. Length
2. Audience awareness
3. Organization
4. Language usage (vocabulary and grammar)

Table 2.4 Unassisted writing criteria categories

1. Clarity and development of ideas
2. Length
3. Organization
4. Language usage (vocabulary and grammar)

For each of the three writing types (i.e., portfolio letter, multi-draft essay, unassisted writing), the number of ratings in which each criterion was considered in arriving at a scoring decision or in explaining assigned scores was tallied. Given the fact that there were a total of twenty ratings (four raters and five portfolios) for each of the three writing types, the total number of times a single criterion could be counted was twenty. The first set of frequency counts reported in the following section indicates the number of ratings, out of a possible twenty, in which each criterion was considered. Thus, this data is indicative of the frequency with which the raters attended to each of the various criteria in judging the quality of the three different types of writing samples.

For a second type of frequency analysis, the transcripts were examined in order to determine the criteria according to which the *marginal* and *unacceptable* writing samples were judged to be deficient. Whereas the initial frequency count was thought to potentially provide

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

information about the criteria raters considered, this second analysis was thought to be more indicative of each criterion's influence in actually determining final scores. Deficiencies were identified as any writing traits which were deemed to not meet the program's standards for an *acceptable* score, according to the rater's interpretation of the scoring rubric descriptors for *acceptable* writing samples. These deficiencies were noted for each *marginal* and *unacceptable* writing sample, and frequency counts were then calculated for each of the three writing types, in a similar fashion to that of the frequency analysis described above.

In addition to the frequency analyses of the criteria used to arrive at scores for the portfolio items, a content analysis of the raters' transcripts was performed in order to investigate any effects that the uniqueness of the portfolio assessment process had on scoring outcomes. This content analysis involved an examination of the transcripts on a rater-by-rater basis as well as on a portfolio-by-portfolio basis in order to identify the ways in which the scoring procedure itself, as described in the background section of this paper, was carried out by the different raters, and how this behavior may have had an influence on scoring decisions.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

3. RESULTS

3.1 Research Question #1: The influence of various criteria on holistic scores

Tables 3.1-3.3 indicate the number of ratings in which a given criterion was considered when deciding final holistic scores for each of the three writing types (i.e., portfolio letter, multi-draft essay, and unassisted writing). The frequency with which a criterion was taken into account for a given writing type is considered to be representative of that criterion's relative importance to the raters in determining scores for that writing type.

Tables 3.1-3.3: Frequency of criteria considered for ratings of each writing type (R = Rater)

Table 3.1 Frequency of criteria considered in rating portfolio letters

CRITERIA	R1	R2	R3	R4	TOTAL
1. Awareness of self as a writer (reflection) or clarity of voice	5	4	5	5	**19**
2. Explanation of assignments & contents of portfolio	5	2	3	5	**15**
3. Language usage (vocabulary and grammar)	1	1	4	3	**9**
4. Organization	1	1	2	0	**4**
5. Length	0	2	1	0	**3**

No specific criteria could be identified in 1 out of the 20 portfolio letter ratings.

Table 3.2 Frequency of criteria considered in rating multi-draft essays

CRITERIA	R1	R2	R3	R4	TOTAL
1. Content*	(5)	(4)	(4)	(5)	**(18)**
1a. Clarity, development, and support of ideas	3	4	1	5	**13**
1b. Clarity of focus	3	0	3	2	**8**
1c. Unity of the essay	1	0	1	2	**4**
1d. Length	0	0	1	0	**1**
2. Organization	4	5	3	4	**16**
3. Language usage (vocabulary and grammar)	3	3	1	3	**10**
4. Audience awareness	2	3	0	0	**5**

Specific criteria were identified in all 20 of the multi-draft essay ratings.

** The frequency counts in parentheses for the umbrella category of "content" represent the number of ratings in which any one of the "content" subcategories was considered.*

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

Table 3.3 Frequency of criteria considered in rating unassisted writings

CRITERIA	R1	R2	R3	R4	TOTAL
1. Clarity and development of ideas	5	3	2	5	**15**
2. Language usage (vocabulary and grammar)	2	2	4	3	**11**
3. Length	1	1	3	1	**6**
4. Organization	1	0	1	2	**4**

No specific criteria could be identified in 3 of the 20 unassisted writing ratings.

Upon examining the data, it is evident that raters favored certain criteria over others in arriving at their holistic assessments. In fact, certain criteria far outweighed others in the ratings for the portfolio letters and unassisted writings in particular. In the portfolio letter ratings, the criterion of *awareness of self as a writer (reflection) or clarity of voice* was considered in all but one of the portfolio letter ratings, and the criterion of *explanation of assignments and contents of portfolio* was considered in fifteen out of the twenty ratings. In contrast, the three remaining portfolio letter criteria of *language usage*, *organization*, and *length* were considered in fewer than half of the ratings for that writing type. A similar finding is evident in the data for the unassisted writings, in which the criteria of *clarity and development of ideas* and *language usage* were considered much more frequently (in fifteen and eleven ratings, respectively) than *length* and *organization* (six and four ratings, respectively) in assessing those writing samples. As for the multi-draft essays, *organization* was the most influential trait, followed by the "content" subcategory of *clarity, development, and support of ideas*, and then *language usage*.

The results also indicate that the relative importance of the criteria categories of *language usage* (vocabulary and grammar) and *organization* may vary. In two out of the three writing types—portfolio letters and unassisted writings—*language usage* was a much greater consideration than *organization*. This trend was most evident for the unassisted writings, in which *language usage* was mentioned as a consideration in eleven of the ratings and *organization* was considered in only four of the ratings. In fact, one rater (R2) did not refer to *organization* in judging any of the unassisted writing samples. As for the portfolio letters, while the criterion of *language usage* was also considered more often than *organization* overall, the data reveal that this was not the case for each individual rater. Raters 1 and 2 both considered *language usage* and *organization* an equal number of times when judging the quality of the portfolio letters. Raters 3 and 4 considered *language usage* more frequently than *organization* in rating the portfolio letters. However, these results did not hold true for the multi-draft essay ratings, in which *organization* was considered more often than *language usage*. This suggests that the relative importance of these different

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

criteria may vary depending on the writing type being assessed, which is not entirely surprising, considering the differences in nature of the three writing types.

These findings regarding the greater influence of *language usage* in comparison to *organization* for the portfolio letters and unassisted writings support those of Sweedler-Brown (1993), whose study found grammar to be more influential than organization in holistic assessments of second language writing. On the other hand, the results of the ratings for the multi-draft essays are more congruent with the findings reported by Chiang (1999) and O'Loughlin (1994), in that they reflect a greater influence of the criterion of *organization* in comparison to *language usage*. Chiang's criteria category of *coherence*, which is to a certain extent inclusive of the current study's categories of *content* and *organization*, was more influential than the categories of *morphology* and *syntax*, which along with vocabulary form this study's category of *language usage*. Similarly, O'Loughlin's category of *organization* was more influential than the category of *grammar & cohesion*, which is somewhat related to the present study's *language usage* criterion.

The results of the main study, as reported here, are generally reflective of and supported by those of the small-scale pilot study in which the same four participants took part two weeks prior to the data collection for the main study. The results of the pilot, which are reported in Appendix D, reflect the relatively strong influence of *awareness of self as a writer* and *explanation of assignments* for the portfolio letters as well as *clarity and development of ideas* for the unassisted writings. They also indicate relatively equal degrees of influence for the three multi-draft essay criteria of *clarity and development of ideas*, *organization*, and *language usage*. Finally, the pilot also provided evidence of the varying degrees of influence of *organization* and *language usage*, depending on writing type.

Whereas the results reported above in Tables 3.1-3.3 indicate the total number of ratings in which each criterion was considered when scoring writing samples, they do not necessarily provide information about exactly how each criterion influenced the actual given scores. In other words, the fact that the quality of a writing sample was discussed in terms of a given criterion does not necessarily mean that the criterion in question played a significant role in determining the final assigned score. Therefore, in order to gain more insight into this question, it is important to look at how perceived writing deficiencies according to these criteria influenced the decisions to find writing samples *marginal* or *unacceptable*. Tables 3.4-3.6 indicate the number of ratings in which *marginal* and *unacceptable* writing samples were deemed to be deficient according to each rating criterion.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

Tables 3.4-3.6: Frequency of writing deficiencies by criteria category for *marginal* and *unacceptable* ratings in each writing type

Table 3.4 Frequency of writing deficiencies by criteria category for *marginal* and *unacceptable* portfolio letter ratings (total = 13)

Criteria	Frequency
1. Awareness of self as a writer (reflection) or clarity of voice	11
2. Explanation of assignments & content of the portfolio	7
3. Language usage (vocabulary and grammar)	5
4. Length	3
5. Organization	2

Table 3.5 Frequency of writing deficiencies by criteria category for *marginal* and *unacceptable* multi-draft essay ratings (total = 12)

Criteria	Frequency
1. Content*	(10)
1a. Clarity, development, and support of ideas	6
1b. Clarity of focus	5
1c. Unity of the essay	4
1d. Length	1
2. Organization	8
3. Language usage (vocabulary and grammar)	3
4. Audience awareness	0

*The frequency count in parentheses for the umbrella category of "content" represents the number of ratings in which *marginal* and *unacceptable* multi-draft essays were deemed to be deficient in any one of the "content" subcategories.

Table 3.6 Frequency of writing deficiencies by criteria category for *marginal* and *unacceptable* unassisted writing ratings (total = 14)

Criteria	Frequency
1. Clarity and development of ideas	9
2. Language usage (vocabulary and grammar)	6
3. Organization	2
4. Length	0

This frequency analysis of writing deficiencies for *marginal* and *unacceptable* writing samples provides further evidence of the relatively high influence of *awareness of self as a writer (reflection)* and *explanation of assignments* on the ratings of the portfolio letters. It also underscores the influence of *clarity and development of ideas* and *language usage* on the ratings of the unassisted

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

writings as well as *clarity, development, and support of ideas* and *organization* on the multi-draft essay assessments. Furthermore, the results of this second analysis reflect the previous findings that *language usage* was more influential than *organization* in determining the scores for the portfolio letters and unassisted writings while *organization* was more influential than *language usage* for the multi-draft essays. With respect to the generalizability of these results, it must be acknowledged that these findings may have been primarily a result of the quality of the portfolios selected, and not necessarily a result of the different degrees of influence of the various rating criteria.

3.2 Research question #2: How raters' perceptions of the relative importance of various criteria match the use of those criteria in actual ratings

In order to answer this research question, the results from the preceding section must be compared with the results of the rater survey on the relative importance of the rating criteria. The results of the criteria rankings from the rater questionnaire are reported below in Tables 3.7-3.9.

Tables 3.7-3.9: Raters' perceptions of relative importance of rating criteria (Source: rater questionnaire)

Table 3.7 Relative importance of rating criteria for portfolio letters
Scale: 5 = most important, 1 = least important

Criteria	Mean	Median	S.D.
1. Awareness of self as a writer (reflection) or clarity of voice	4.33	5	.90
2. Explanation of assignments and contents of portfolio	4.27	5	1.10
3. Language usage (vocabulary and grammar)	3.67	4	.98
4. Organization	3.27	3	1.10
5. Length	2.20	2	1.08

Table 3.8 Relative importance of rating criteria for multi-draft essays
Scale: 4 = most important, 1 = least important

Criteria	Mean	Median	S.D.
1. Content	3.86	4	.36
2. Organization	3.43	3	.51
3. Audience awareness	2.38	2	.96
4. Language usage (vocabulary and grammar)	2.36	2	.84

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

Table 3.9 Relative importance of rating criteria for unassisted writings
Scale: 4 = most important, 1 = least important

Criteria	Mean	Median	S.D.
1. Clarity and development of ideas	3.87	4	.35
2. Language usage (vocabulary and grammar)	2.87	3	.64
3. Length	2.47	3	.83
4. Organization	2.33	3	.98

In comparing the results illustrated in the tables above with those reported in the previous section (Tables 3.1-3.3 and Tables 3.4-3.6), it is clear that the rating behavior of the participants in the verbal report phase of the study closely matched the program's overall perceptions of the relative importance of the various rating criteria. In fact, the results of these rankings almost perfectly mirror the results of the verbal report, as previously illustrated in the frequency analysis tables. As in the verbal report data, the criterion of *awareness of self as a writer (reflection)* for the portfolio letters, the umbrella category of *content* for the multi-draft essays, and the criterion of *clarity and development of ideas* for the unassisted writings were the most important criteria for their respective writing types. This result is evident not only in the mean and median figures, but also in the fact that the highest-ranked criterion for each of the three writing types resulted in the lowest standard-deviation figures, which suggests that the raters agreed most on the importance of the highest-ranked criterion for each writing type. In addition, these rankings further support the previously noted tendency for the importance of *organization* and *language usage* to vary according to the writing type being assessed. Once again, *language usage* was deemed to be more important than *organization* in assessing the portfolio letters and unassisted writings, while the opposite was true in the case of the multi-draft essays.

3.3 Research question #3: The effect of the portfolio assessment process on scoring outcomes

As described previously, this program's procedure for arriving at the overall score for a given portfolio consists of a set formula which takes into account the individual scores assigned to each of the three writing samples in the portfolio. In essence, this procedure can be described as bottom-up, in that the overall score is based on the three scores of the portfolio's constituents. Without question, this bottom-up scoring procedure was by far most influential in determining final portfolio grades. However, the content analysis of the rater transcripts identified what appears to be a second, more top-down assessment process employed by raters. Whereas the bottom-up process is more analytic in determining portfolio scores, this second process is more holistic in that

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

a rater considers the quality of the portfolio as a whole entity, as opposed to a sum of its parts. Moreover, there is evidence to suggest that this top-down process can also be influential in determining the final portfolio scores, although this is not necessarily the case for all raters.

Perhaps the most significant evidence of this top-down process is that which illustrates how individual writing sample scores can be influenced by the overall perception of a given portfolio's quality. The following excerpt from Rater 2's transcript reflects this type of rater behavior. Here Rater 2 has just finished reading the unassisted writing sample of Portfolio C.

> *Hmm, well the unassisted writing is clearly* ***marginal****, and I think that's what this portfolio probably is. I think then I'm gonna marginalize the letter, and I'm marginalizing the letter because of the organization of that third paragraph, which is* ***really*** (emphasis hers) *confusing to me, um, because it flips back and forth.*

From this segment of the transcript, it is clear that the rater has formed an overall opinion on the quality of the entire portfolio after having read the third and final writing sample, the unassisted writing. She then goes back to the portfolio letter, which was the first item she read, and assigns it a score of *marginal*. This is significant because it ultimately has an effect on the final grade assigned to the portfolio. Initially, when she first read the letter, she tentatively decided to score it as *unacceptable*, which would have resulted in an *unacceptable* grade for the entire portfolio, based on the scoring formula. However, after arriving at a holistic impression of the overall *marginal* quality of the portfolio, she decides to change the individual score of the unassisted writing to reflect that top-down impression, which ultimately leads to a final score of *marginal* for the portfolio as a whole.

A similar example of the potential effect of this top-down rating process on scoring outcomes is found in Rater 3's verbal report. In this segment, the rater has begun to consider a score for the multi-draft essay in Portfolio E.

> *Well, this (essay) is* ***marginal*** *at worst, not* ***unacceptable****. The question is, is it* ***acceptable****? Shoot. It's basically very simple. Mmm. My feeling is that I can't let this be, in terms of the whole portfolio, I can't let this be an acceptable portfolio. Ah, the letter was too close to* ***marginal****—I'm sorry—too close to* ***unacceptable,*** *and this is too weak.*

After reporting this, the rater reexamines the essay, seemingly with the intent to find traits that would justify assigning a score of *marginal*. Eventually, he does in fact assign a *marginal* score, despite the fact that he admits possibly being a little harsh in doing so. What is important is that this score for the individual writing sample seems to be at least indirectly influenced by the rater's overall impression of the less-than-acceptable quality of the portfolio as a whole. By assigning a score of *marginal* to the multi-draft essay, he ensures that the portfolio will receive a *marginal* score at best, thus affirming his top-down assessment of its quality. Unlike the previous example, it is

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

significant to note that in this case, the rater has arrived at his top-down assessment without having read the unassisted writing sample, which is the third and final component of the portfolio. This finding, like those reported by Hamp-Lyons and Condon (1993), indicates that readers' assessments may be rendered before the entire portfolio is read, which is a severe threat to the assessment's validity.

Although these examples indicate some degree of influence of this top-down assessment process, it was definitely not as influential as the bottom-up process involved in the program's portfolio scoring formula. In the majority of the assessments, the raters relied heavily on the individual scores they assigned in the bottom-up assessment process. Rater 1 in particular seemed to be completely resistant to any effects of a top-down assessment. The following comment, which she makes after finishing her assessment of Portfolio C, provides evidence to support this claim.

> *So, I guess, um, well, I don't know if I feel* ***totally*** (emphasis hers) *comfortable giving this whole portfolio an [acceptable], but, um, that's what the final grades come out to be.*

Despite having assigned scores of *acceptable* to each of the three writing samples in this portfolio, the rater has misgivings concerning whether or not the portfolio as a whole merits an *acceptable* score. As a result, her top-down assessment, which questions the portfolio's quality, is in conflict with her bottom-up assessment. Nonetheless, she disregards her doubts about the portfolio's overall quality and instead relies on the bottom-up assessment. Her decision not to act upon her top-down assessment is significant because she was the only rater to find that particular portfolio *acceptable*; all three of the other raters found it to be *marginal*.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

4. Discussion

4.1 Interpretation of results and implications for portfolio assessment

With regard to the relative influence of various criteria on holistic scores, it is obvious from the results of this study that raters are more affected by certain factors than others. This is not to be unexpected since some characteristics of writing are certainly more important than others. For example, it would be somewhat absurd to suggest that the number of words in an essay should be equally important as the ability of the writer to express his or her ideas clearly. However, with the goal of holistic scoring being that of assessing overall writing ability, it would seem to be in the best interest of the assessment to ensure that raters are influenced by all of the intended criteria to a reasonable extent. Then and only then would the assessment accurately reflect the construct of overall writing proficiency, and not writing proficiency according to a single, excessively influential trait.

That being said, the findings of this research also indicate that the degree of influence of given criteria is to a certain extent dependent on the writing type being assessed. In this study, for example, *organization* was a much more influential criterion in the multi-draft essay assessments than in the portfolio letter and unassisted writing assessments. Perhaps this can be explained by assuming that raters expect more in terms of organization in a multi-draft essay than in an unassisted writing. It is possible that raters see organization as something that develops throughout the process of writing a more substantial piece, such as a multiple-draft essay, as opposed to an item in which the writer invests less time and effort, such as an unassisted writing. Furthermore, the multiple-draft essay is a task on which the writer receives feedback and guidance from the instructor. For this reason, raters may expect better organization for this writing type, considering the fact that the writer has the opportunity to incorporate teacher suggestions directed at organization. Regardless of the reasoning behind these expectations, it is nonetheless apparent that raters are more influenced by different criteria according to task type. Therefore, at least in the context of this particular portfolio assessment program, holistic impressions of writing quality may vary significantly according to the specific writing task being assessed. In other words, the construct of writing proficiency is not static across genres of writing in the case of a multidimensional portfolio assessment.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

These findings raise concerns about whether holistic scoring is actually the most valid scoring procedure for the variety of writing types that are included in portfolios such as these. While raters may lend more consideration to the full set of rating criteria for certain writing types, such as the multi-draft essays in this study, they may be more heavily influenced by one or two specific criteria on other writing types. Perhaps then it would be in a program's best interests to implement a primary-trait scoring procedure for the assessments of those writing types that are most heavily influenced by a single trait. For example, if a program decides that the quality of a portfolio letter is best determined by the quality of the writer's ability to reflect on his or her writing experience, then it might be best to judge the letters on that characteristic alone. Certainly this would make the practice of setting standards much more feasible. Instead of setting standards for four or five different criteria, the program could focus on setting standards for a single criterion, which is without a doubt a much easier and much more practical chore. Then, by better ensuring that raters are focusing on the same criterion and by devoting attention to defining clearer standards for that criterion, the assessment would become more reliable, which would in turn increase the assessment's validity to a certain extent. There is, however, a caveat that must be acknowledged for any primary-trait assessment, which is the possibility that readers may very well be unable to ignore the other features of writing and judge the sample only on the basis of the primary trait (Hamp-Lyons, 1991).

On the other hand, it might be in a program's best interests to ensure that a wide array of criteria are considered. If, for example, it is indeed the case that organization is an important aspect of a portfolio letter, then steps should be taken in order to guarantee that raters consider it to a reasonable extent. However, judging by the results of this study, it is not clear whether a holistic rating procedure can provide such a guarantee. It is simply very easy to be overly influenced by salient strengths or shortcomings in the quality of the writing according to one or two criteria. This is especially true in the case of assessing nonnative writers, whose development as writers progresses unevenly across different skill domains such as grammatical control, organization, expression of ideas, etc. (Hamp-Lyons, 1991). For this reason, in situations where a variety of criteria are deemed to be important, it might be beneficial to adopt a multi-trait scoring procedure in which each sample of writing receives a different score for each individual criterion. This would not necessarily entail using a prescribed formula of combining those scores to reach an overall score for the writing sample, but it might at least encourage raters to consider the full set of criteria in each of their assessments before assigning final scores. As a consequence of encouraging the full consideration of

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

all criteria, the assessment would become more holistic in nature and thus a more valid measure of overall writing proficiency.

The effect of the portfolio assessment process on the scoring outcomes must also be addressed when considering the validity of the scoring procedure. The raters' use of a top-down assessment process in conjunction with a bottom-up process in this study underscores the complexity of this form of assessment. While it may be true, as Hamp-Lyons and Condon (1993) point out, that readers are most likely to assess the component texts individually and weigh each of them in light of the others when judging a portfolio's overall quality, it is also the case that readers may decide on a score for a particular component text based in part on the overall quality of the portfolio as a whole. At least this may be true in programs such as this where scores are assigned to all of the individual writing samples in a portfolio, not just the portfolio as a whole.

The implications of this "two-directional" process of assessment in terms of being a benefit or drawback are not necessarily obvious. For the most part, it would seem advantageous for the assessment process to occur in both directions (i.e., top-down and bottom-up). In a sense, this might serve as a system of checks and balances in which the top-down assessment could either reinforce or contradict the bottom-up assessment. In the latter case, the rater would ideally be encouraged to reconsider the portfolio's quality in light of such a conflict. In fact, for this very reason, readers in this particular program were actually encouraged to consider the quality of the portfolio as a whole after scoring all of the individual writings in it. However, the results of this study indicate that raters do not engage in this process consistently, nor are they influenced by the process equally. This lack of consistency highlights another complication arising within the context of portfolio assessment. When the different raters behave inconsistently, the interrater reliability of the assessment is likely to suffer. Therefore, although a two-directional process of portfolio assessment may be a benefit in that it provides two informative perspectives, it may damage the measure's reliability if all raters do not engage in the process and are not equally influenced by it. Moreover, the evidence cited above indicating that raters may arrive at their top-down assessments before considering all items in a portfolio is equally troublesome. The top-down perspective is only meaningful if it is inclusive of all of the work inside the portfolio.

4.2 Limitations of the study

The most obvious limitations of this study are perhaps those related to the sample characteristics, which contribute to the limited generalizability of the study's findings. While the survey phase of the study included a sample of participants which could be described as

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

representative of the program as a whole, the verbal report phase was quite limited in that there were only four participants. Considering the fact that ten or more raters were typically involved in any given portfolio assessment session in this program, data collected from four participants may not fully reflect the rating behavior of the program as a whole. In addition to the limited number of participants, the sample size of the portfolios that were read in the study was quite small, and these portfolios were all from students in the same composition class. For these reasons, the portfolios in the study did not completely represent the variety of writing that would have been typically found in the portfolios of an actual assessment. As a result of these limitations, it would be difficult to defend any sweeping generalizations based on the study's findings that would hold true in all instances of portfolio assessment in the context of this particular program or others.

There are also limitations associated with the method of data collection. As is the case with all methodologies that rely on verbal report, it is impossible to assume that all of the participants' thought processes surfaced in the data. For this reason, it is quite possible that the different criteria were actually more or less influential on the raters' judgments than they appear to be upon examination of the results of the study. For example, while organization was mentioned as a consideration in only four of the twenty ratings of unassisted writings, it is possible that it was actually considered more often than that number indicates. The raters may have been aware of all of the criteria but attended to some of them less than others in certain situations and were thus less likely to report those considerations verbally.

Finally, there are limitations regarding the analysis of the data. Although the categorization of rater comments in the verbal report was fairly straightforward in most cases, there were instances in which problems arose in judging the category to which a comment most pertained. This was particularly true in attempting to distinguish some comments that seemed to address issues related to the multi-draft essay criteria categories of *content*, *organization*, or both. The division between these two criteria categories was not always clear, so the researcher was forced to interpret certain problematic comments in terms of which of these criteria categories they most pertained to. A fact which further clouds this issue is that others in the fields of second language acquisition, pedagogy, and assessment would undoubtedly disagree with the some of the distinctions this particular program made in defining each of the criteria categories. For example, the criteria category of *organization* included the extent to which the writing progresses fluently from one section to the next. It is quite possible that others would decide to associate this trait with the category of *content* since it is related to the writer's ability to express his or her ideas coherently. These kinds of potential disagreement stem from the fact that practitioners in the field of writing assessment are

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

prone to relying on criteria categories which are identified with vague terms such as *content*. In order for the results of research in the field to be meaningful and comparable, it is necessary to clarify exactly which writing traits these broad categories represent, as this study has attempted to do in identifying the various traits associated with the category of *content*.

4.3 Suggestions for future research

Providing evidence of the extent to which raters are influenced by various scoring criteria is only one aspect of a full investigation into the validity of any scoring procedure in portfolio assessment, be it holistic or otherwise. Another very important area for future research would be to examine the standards of quality that raters use when assessing portfolios. Once the criteria used by raters have been determined, it is necessary to investigate exactly what constitutes varying levels of quality according to those criteria. To put it in terms of the portfolio assessment program described in this study, what is the difference, for example, between *acceptable*, *marginal*, and *unacceptable* organization? What standards do raters compare samples of writing against in order to determine their quality?

This issue relates very closely to two very important components of the portfolio assessment process identified earlier in this paper: the standardization meeting and the scoring rubric. As Condon and Hamp-Lyons (1993) point out, regularly conducted standardization meetings are an integral part of any portfolio assessment program since they serve to define standards and build consensus among the raters on how these standards should be applied in determining the quality of writing in a portfolio. However, judging by the data collected in this study, it is not apparent whether raters in fact base their decisions on the standards that are set in these meetings. Out of the sixty ratings conducted in this study, specific mention to a standard discussed in such a meeting was made only one time. This raises serious concerns about whether raters are in fact applying the standards set in these meetings, or if they are instead applying standards based on their own personal beliefs of what constitutes different degrees of writing quality, which is an issue echoed by other researchers (Charney, 1984; Reed & Cohen, 2001).

An investigation into the application of standards would also undoubtedly need to address issues related to the scoring rubric. While the standardization meetings ideally define how the language of the descriptors on the rubric should be interpreted by the raters, it is important to identify how raters in fact interpret such rubric language in actual rating situations when making judgments of writing quality. To refer back to a previous example, where do raters draw the line between *very clear organization* and *somewhat clear organization*? Research into this type of

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

question could compare interpretations of standards of quality across raters in order to determine the extent to which they agree on such vague distinctions made in the rubric descriptors. Furthermore, this line of research could identify particularly problematic areas in the raters' interpretations of standards in order to further the objective of improving the consistency of those interpretations (i.e., interrater reliability).

4.4 Pedagogical implications

A very important dimension of validity has been all but ignored up to this point in this paper. While previously discussed issues have focused mainly on construct validity, it is also important to consider the consequential validity of a performance assessment such as portfolios. Consequential validity refers to the ways in which the assessment influences the instructional practices in the program in which it is implemented. Messick (1989) classifies this as a second type of validity and claims that it must be addressed in conjunction with evidential validity (e.g., construct validity). Similarly, Linn et al. (1991) identify the consequences of the assessment as one of eight validation criteria for any performance-based assessment.

Any discussion of the pedagogical implications of the results of research into a form of language assessment inevitably focuses on the washback effect of the assessment, or in other words, its effect on teaching and learning. Ideally, an assessment should accurately reflect and positively influence the practice of teaching, just as pedagogy should reflect and influence language assessment. An assessment that is beneficial to a language program is one which promotes positive changes in the curriculum and actual classroom instruction in that program. When this occurs, the assessment can be justified as possessing consequential validity.

The results of this study alone can not provide empirical evidence of the consequential validity of portfolio assessment. Further research would be required in order to investigate whether classroom instruction in fact reflects the findings of this study and whether such classroom practices have been positively influenced by the portfolio assessment. Specifically, it would be necessary to determine if instructors teach writing in a way that is in agreement with the relative influences of the various assessment criteria in judging the quality of writing. For this particular portfolio assessment program to possess consequential validity, the instructional emphasis in teaching students to write multi-draft essays, for example, would have to be on expressing ideas with adequate clarity, development, support, and organization. In theory, this assessment would seem to have some degree of consequential validity since the goals of communicative language teaching are weighted heavily in favor of expression of ideas, and that emphasis is reflected to some extent in

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

the rating behavior of the participants in this study. Therefore, it would be plausible to assume that the portfolio assessment would beneficially influence the practice of teaching.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

5. Conclusion

Research of this type into second language writing portfolio assessment remains somewhat limited, and until more serious inquiries into issues related to validity and reliability are conducted, the value of this form of writing assessment will remain in question. The research agenda outlined by Hamp-Lyons and Condon (2000) is a potentially useful beginning to this line of research. While the advantages of portfolio assessment are difficult to deny, the possible shortcomings are equally clear. Unless research and practice can significantly reduce the threats to the validity and reliability of portfolio assessment, it can not ethically be considered as a measure upon which to base high-stakes decisions regarding the writing proficiency of students of second and multiple languages. However, this does not necessarily mean that portfolio assessment is a lost cause. As documented in the literature reviewed in this paper, no form of writing assessment has been found to be perfect, and yet the traditional forms of assessment still exist in many contexts and continue to provide potentially useful information about the subjects of those assessments. Therefore, the best course of action is to base overall judgments of writing proficiency on the results of multiple measures, which may include a portfolio assessment instrument.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

References

Anastasi, A. (1976). *Psychological testing* (4th ed.). New York: Macmillan.

Armstrong Smith, C. (1991). Writing without testing. In P. Belanoff & M. Dickson (Eds.), *Portfolios: Process and product* (pp. 279-292). Portsmouth, NH: Boynton/Cook.

Broad, R. L. (1994). "Portfolio scoring": A contradiction in terms. In L. Black, D. A. Daiker, J. Sommers, & G. Stygall (Eds.), *New directions in portfolio assessment* (pp. 263-276). Portsmouth, NH: Boynton/Cook.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly, 32,* (4), 653-675.

Byrd, P., & Nelson, G. (1995). NNS performance on writing proficiency exams: Focus on students who failed. *Journal of Second Language Writing, 4,* (3), 273-285.

Callahan, S. (1995). Portfolio expectations: Possibilities and limits. *Assessing Writing, 2,* (2), 117-151.

Camp, R. (1993). The place of portfolios in our changing views of writing assessment. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 183-212). Hillsdale, NJ: Lawrence Erlbaum Associates.

Camp, R., & Levine, D. S. (1991). Portfolios evolving: Background and variations in sixth- through twelfth-grade classrooms. In P. Belanoff & M. Dickson (Eds.), *Portfolios: Process and product* (pp. 194-205). Portsmouth, NH: Boynton/Cook.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18,* (1), 65-81.

Chiang, S. Y. (1999). Assessing grammatical and textual features in L2 writing samples: The case of French as a foreign language. *The Modern Language Journal, 83,* (2), 219-232.

Cohen, A. D. (1994). *Assessing language ability in the classroom* (2nd ed.). Boston: Heinle & Heinle.

Condon, W., & Hamp-Lyons, L. (1991). Introducing a portfolio-based writing assessment: Progress through problems. In P. Belanoff & M. Dickson (Eds.), *Portfolios: Process and product* (pp. 231-247). Portsmouth, NH: Boynton/Cook.

Conner-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly, 29,* (4), 762-765.

Elbow, P. (1991). Foreword. In P. Belanoff & M. Dickson (Eds.), *Portfolios: Process and product* (pp. ix-xvi). Portsmouth, NH: Boynton/Cook.

Elbow, P., & Belanoff, P. (1986). State University of New York, Stony Brook, portfolio-based evaluation program. In P. Connolly & T. Vilardi (Eds.), *New methods in college writing programs: Theories in practice* (pp. 95-105). New York: MLA.

Gitomer, D. H. (1993). Performance assessment and educational measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 241-263). Hillsdale, NJ: Lawrence Erlbaum Associates.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69-87). Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Norwood, NJ: Ablex.

Hamp-Lyons, L. (1995a, March). "Portfolios with ESL writers: What the research shows." Paper presented at the 29th annual TESOL Convention, Long Beach, CA.

Hamp-Lyons, L. (1995b). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly, 29,* (4), 759-762.

Hamp-Lyons, L. (1996a). Applying ethical standards to portfolio assessment of writing in English as a second language. In M. Milanovich & N. Saville (Eds.), *Performance testing and assessment: Selected papers from the 15th Language Testing Research Colloquium* (pp. 151-164). Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1996b). The challenges of second-language writing assessment. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 226-240). New York: MLA.

Hamp-Lyons, L., & Condon, W. (1993). Questioning assumptions about portfolio-based assessment. *College Composition and Communication, 44*, (2), 176-190.

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

Hamp-Lyons, L., & Kroll, B. (1996). Issues in ESL writing assessment: An overview. *College ESL, 6,* (1), 52-72.

Horowitz, D. (1991). ESL writing assessments: Contradictions and resolutions. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 71-85). Norwood, NJ: Ablex.

Huerta-Macias, A. (1995). Alternative assessment: Responses to commonly asked questions. *TESOL Journal, 5*, 8-11.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41,* (2), 201-213.

Huot, B. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206-236). Cresskill, NJ: Hampton Press.

Huot, B. (1994). Beyond the classroom: Using portfolios to assess writing. In L. Black, D. A. Daiker, J. Sommers, & G. Stygall (Eds.), *New directions in portfolio assessment* (pp. 325-333). Portsmouth, NH: Boynton/Cook.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*, 15-21.

Lyman, H. B. (1978). *Test scores and what they mean* (3rd ed.). Englewood Cliffs: Prentice

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*, 5-11.

Miller, M. D., & Legg, S. M. (1993). Alternative assessment in a high-stakes environment. *Educational Measurement: Issues and Practice, 12,* (2), 9-15.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

O'Loughlin, K. (1994). The assessment of writing by English and ESL teachers. *Australian Review of Applied Linguistics, 17,* (1), 23-44.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly, 17,* (4), 651-671.

Popham, J. W. (1981). *Modern educational measurement.* Englewood, NJ: Prentice.

Reed, D. J., & Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. In C. Elder et al. (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 82-96). Cambridge: Cambridge University Press.

Roemer, M., Schultz, L. M., & Durst, R. K. (1991). Portfolios and the process of change. *College Composition and Communication, 42,* (4), 455-469.

Ruetten, M. (1994). Evaluating ESL students' performance on proficiency exams. *Journal of Second Language Writing, 3,* (2), 85-96.

Shale, D. (1996). Essay reliability: Form and meaning. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 76-96). New York: MLA.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing, 5,* (2), 163-182.

Sweedler-Brown, C. O. (1993). ESL essay evaluation: The influences of sentence-level and rhetorical features. *Journal of Second Language Writing, 2,* (1), 3-17.

Vaughan, C. (1991). Holistic assessment: What goes on in the raters' minds? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-125). Norwood, NJ: Ablex

White, E. M. (1985). *Teaching and assessing writing.* San Francisco: Jossey-Bass.

White, E. M. (1994). Portfolios as an assessment concept. In L. Black, D. A. Daiker, J. Sommers, & G. Stygall (Eds.), *New directions in portfolio assessment* (pp. 25-39). Portsmouth, NH: Boynton/Cook.

Wiggins, G. (1994). The constant danger of sacrificing validity to reliability: Making writing assessment serve writers. *Assessing Writing, 1*, 129-139.

Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of assessment. In G. Grand (Ed.), *Review of Research in Education, Vol. 17.* Washington, DC: American Educational Research Association.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

APPENDIX A: PORTFOLIO SCORING RUBRIC

	Acceptable (= A)	Marginal (= M)	Unacceptable (= U)
Portfolio Letter Grade (circle): A M U	Clearly explains assignments & content of portfolio; shows awareness of self as a writer (reflection) or clear voice; 300-500 words. Organization & language satisfactory (see criteria for essays in Marginal column).	Explains assignments and content of portfolio, but incompletely and/or unclearly and/or with insufficient detail; shows minimal awareness of self as a writer (reflection) or minimal evidence of voice. Length may be inappropriate for task. Organization & language satisfactory (see criteria for essays below).	Doesn't explain assignments or contents; doesn't show self-awareness (reflection) or evidence of voice; may be less than 200 words. May have inadequate range and/or control of grammar and/or vocabulary. May have poor organization. Problems may interfere with meaning.
Essay Grade (circle): A M U	Topic or title:		
Content	Ideas are clear, developed and well-supported. Focus is clear. Essay is unified. Generally at least 600 words.	Ideas need more support, development, and/or clarity. Focus is somewhat unclear. There may be some unity problems.	Ideas are mostly unsupported, undeveloped, and/or unclear. Focus is unclear. There may be some unity problems. May appear plagiarized.
Audience Awareness	Shows awareness of audience; e.g., articles used as the basis for the essay are explained to an outside reader.	Shows only limited awareness of audience; e.g., articles used as the basis for the essay are not adequately explained to the reader.	Doesn't show awareness of audience; e.g., articles used as the basis for the essay are not explained to the reader.
Organization	Clear, with logical and fluent progression from one part of the essay to the next.	Mostly clear but some confusion in organization and/or somewhat illogical; some ideas do not progress fluently from one part to the next.	Unclear and/or illogical; ideas do not progress fluently from one part of the essay to the next.
Language Usage	Adequate range and control of vocabulary and grammar.	Some problems with range and/or control of grammar and/or vocabulary, but problems generally do not interfere with meaning.	Inadequate range and/or control of grammar and/or vocabulary; problems may interfere with meaning.
Unassisted Writing Grade (circle): A M U	Topic or title:		
	Ideas are clear, with few lapses; some development of ideas. Length satisfactory for the task (usually at least 300 words). Organization & language satisfactory.	Ideas are clear, with some lapses; may be too short (possibly less than 300 words). Organization and language are mostly satisfactory.	Ideas are unclear; may be too short; little or no development. Organization and language may be weak.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

Appendix B: Rater Questionnaire

Part I. Background Information

Name: _____

Age: 20-30 31-40 41-50 over 50

Gender: Male Female

Estimated number of years of adult ESL teaching experience

_____1-5 years
_____6-10 years
_____11-15 years
_____16-20 years
_____21-25 years
_____more than 25 years

Estimated number of years of experience teaching ESL composition

_____1-5 years
_____6-10 years
_____11-15 years
_____16-20 years
_____21-25 years
_____more than 25 years

Part II. ESL Writing Portfolio Experience

1. How many adult ESL writing portfolio reading/rating sessions have you participated in?

   _____1-5 sessions
   _____6-10 sessions
   _____11-15 sessions
   _____16-20 sessions
   _____21-25 sessions
   _____more than 25 sessions

2. How many, if any, of these rating sessions were in ESL writing portfolio programs other than [this program's]?

3. Please indicate when you last read/rated portfolios in [this program]: _____

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

4. Please indicate the degree to which you agree or disagree with the following statements:

5	4	3	2	1
strongly agree				**strongly disagree**

_____a) Portfolio assessment in general is a **valid** (meaningful) way to evaluate the writing of adult ESL students.

Portfolio assessment in general is a more **valid** (meaningful) way to evaluate the writing of adult ESL students than…

_____ b) timed-essay assessments.

_____ c) final assessments from each individual's writing instructor.

_____ d) Portfolio assessment in general is a **reliable** (fair) way to evaluate the writing of adult ESL students.

Portfolio assessment in general is a more **reliable** (fair) way to evaluate the writing of adult ESL students than…

_____ e) timed-essay assessments.

_____ f) final assessments from each individual's writing instructor.

The remaining questions deal specifically with your experience in [this intensive English program's] current portfolio assessment program. For questions 4-8, please rank the items in each list on the scale provided according to your perception of each item's importance with respect to the other items in the list. ***If you believe that certain items are of equal importance, assign them the same number.***

Sample Responses:

1. Please rank the three portfolio components from 1 to 3 according to your opinion of their importance in determining the portfolio's overall quality.
(1 = least important piece, 3 = most important piece)

__1__ portfolio letter
__3__ multi-draft essay
__2__ unassisted writing

or

__2__ portfolio letter
__2__ multi-draft essay
__1__ unassisted writing

or

__2__ portfolio letter
__1__ multi-draft essay
__1__ unassisted writing

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

A) Please rank the **three portfolio components** from 1 to 3 according to your opinion of their importance in determining the portfolio's overall quality.
(1 = least important piece, 3 = most important piece)

_____ portfolio letter

_____ multi-draft essay

_____ unassisted writing

B) Please rank the **three portfolio components** from 1 to 3 according to your perception of the [program's] opinion of their importance in determining the portfolio's overall quality.
(1 = least important piece, 3 = most important piece)

_____ portfolio letter

_____ multi-draft essay

_____ unassisted writing

5. A) Please rank the following aspects of the **portfolio letter** from 1 to 5 according to your opinion of their importance in determining the letter's overall quality.
(1 = least important aspect, 5 = most important aspect)

_____ explanation of assignments and content of portfolio

_____ awareness of self as a writer (reflection) or clear voice

_____ length of the letter

_____ organization

_____ language usage

B) Please rank the following aspects of the **portfolio letter** from 1 to 5 according to your perception of the [program's] opinion of their importance in determining the letter's overall quality. (1 = least important aspect, 5 = most important aspect)

_____ explanation of assignments and content of portfolio

_____ awareness of self as a writer (reflection) or clear voice

_____ length of the letter

_____ organization

_____ language usage

6. A) Please rank the following aspects of the **multi-draft essay** component of the portfolio from 1 to 4 according to your opinion of their importance in determining the essay's overall quality.
(1 = least important aspect, 4 = most important aspect)

_____ content

_____ audience awareness

_____ organization

_____ language usage

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

B) Please rank the following aspects of the **multi-draft essay** component of the portfolio from 1 to 4 according to your perception of the [program's] opinion of their importance in determining the essay's overall quality. (1 = least important aspect, 4 = most important aspect)

_____ content

_____ audience awareness

_____ organization

_____ language usage

7. A) Please rank the following aspects of the **multi-draft essay's content** from 1 to 4 according to your opinion of their importance in determining the quality of the essay's content. (1 = least important aspect, 4 = most important aspect)

_____ clarity, development, and support of ideas

_____ clarity of essay's focus

_____ unity of the essay

_____ length of the essay

B) Please rank the following aspects of the **multi-draft essay's content** from 1 to 4 according to your perception of the [program's] opinion of their importance in determining the quality of the essay's content. (1 = least important aspect, 4 = most important aspect)

_____ clarity, development, and support of ideas

_____ clarity of essay's focus

_____ unity of the essay

_____ length of the essay

8. A) Please rank the following aspects of the **unassisted writing** component of the portfolio from 1 to 4 according to your opinion of their importance in determining the piece's overall quality. (1 = least important aspect, 4 = most important aspect)

_____ clarity and development of ideas

_____ length of the writing

_____ organization

_____ language usage

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

B) Please rank the following aspects of the **unassisted writing** component of the portfolio from 1 to 4 according to your perception of the [program's] opinion of their importance in determining the piece's overall quality. (1 = least important aspect, 4 = most important aspect)

_____ clarity and development of ideas

_____ length of the writing

_____ organization

_____ language usage

9. In your opinion, which of the following best describes the criteria used on the portfolio scoring rubric?

   a) too broad for an accurate assessment
   b) too specific for an accurate assessment
   c) neither too broad nor too specific

10. With what frequency are you uncertain of the scores (acceptable, marginal, or unacceptable) of each of the three writing components (portfolio letter, multi-draft essay, and unassisted writing) of the portfolios you read?

   a) always
   b) often
   c) sometimes
   d) rarely
   e) never

11. If you indicated any amount of uncertainty on the previous question, to which of these reasons do you most attribute this uncertainty? (If you indicated no uncertainty, disregard this question.)

   a) the language of the descriptors on the scoring rubric (i.e., rubric language is too broad or too specific)
   b) conflicts arising from writing characteristics of differing quality (e.g., clear and well-developed ideas but weak organization and use of language)
   c) other (please explain):

   _____

   _____

   _____

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

APPENDIX C: VERBAL REPORT GUIDELINES

To begin, turn on the recorder at your booth and state your name for identification purposes.

Starting a new text

1. Read the portfolios and their component texts in the order in which they are arranged on your desk.
2. State the designated identification letter of the portfolio you are reading (e.g., *Portfolio A*).
3. State whether the text you are beginning to read is the portfolio letter, multi-draft essay, or unassisted writing.

As you read each text

1. Verbalize as many of your thoughts as possible.
2. Be as specific as possible. Try to provide details about any general comments (e.g., "good organization") by referring to specific features of the writing.
3. Every so often, please give updates on where you are in a given text. Also, make sure you mention if you are skipping ahead, going back, or looking at preliminary drafts.

When assigning a final score to a text

1. State the grade you are giving and mark it on the scoring rubric.
2. Explain why you are assigning that grade, mentioning all factors that influenced your decision. Again, when you make a general comment (e.g., "good organization"), please try to clarify by referring to specific aspects of the writing.
3. For each score you assign, indicate the single most influential aspect of the writing that prompted you to give that particular score.

Upon completion of an entire portfolio

Once you have finished reading and scoring all three texts of a given portfolio, put the completed scoring sheet in the folder with the portfolio texts and move on to the next portfolio.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

APPENDIX D: SUMMARY OF PILOT STUDY DESIGN AND RESULTS

Pilot study design

- **Participants:** the same four raters who participated in the main study
- **Instruments:** verbal report, essentially conducted as described in Sections 3.2.2
- **Data collection:** Raters read two portfolios, providing think-aloud verbal reports as they read and upon considering a score for each writing sample.
- **Data analysis:** Frequency counts of the number of ratings in which each assessment criterion was considered when judging the overall quality of the sample were performed. Total number of ratings for each writing type = 8 (4 raters & 2 portfolios).

Pilot study results

Tables D1-D3: Frequency of criteria considered for ratings of each writing type (R = Rater)

Table D1. Frequency of criteria considered in rating portfolio letters

Criteria	Frequency
1. Awareness of self as a writer (reflection) or clear voice	8
2. Explanation of assignments & content of the portfolio	7
3. Language usage (vocabulary and grammar)	5
4. Length	4
5. Organization	3

Table D2. Frequency of criteria considered in rating multi-draft essays

Criteria	Frequency
1. Content*	(8)
1a. Clarity, development, and support of ideas	6
1b. Clarity of focus	4
1c. Unity of the essay	3
1d. Length	0
2. Organization	5
3. Language usage (vocabulary and grammar)	5
4. Audience awareness	3

*The frequency count in parentheses for the umbrella category of "content" represents the number of ratings in which any one of the "content" subcategories was considered.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/

Table D3. Frequency of criteria considered in rating unassisted writings

Criteria	Frequency
1. Clarity and development of ideas	7
2. Language usage (vocabulary and grammar)	4
3. Length	3
4. Organization	1

- Content was considered most often in determining the quality of the writing samples in the portfolios. This supports the findings of some previous research into single-sample essay assessments (O'Loughlin, 1994; Song & Caruso, 1996; Vaughan, 1991).
- Grammar was considered more often than organization in the ratings for the portfolio letters and unassisted writings. This finding supports those of Sweedler-Brown (1993).
- Language usage and organization were considered to similar extents on the ratings for the multi-draft essays.

© 2013, Regents of the University of Minnesota. Permission to reprint must be sought from the CARLA office. For information see: www.carla.umn.edu. Originally published as: Conrad, C. J. (2001). Second language writing portfolio assessment: The influences of the assessment criteria and the rating process on holistic scores (CARLA Working Paper #20). Minneapolis: University of Minnesota, Center for Advanced Research on Language Acquisition. Retrieved from www.carla.umn.edu/resources/working-papers/