COMPOSITION EVALUATION AS AN UNFOLDING PROCESS:

THE IMPACT OF ESSAY FEATURES ON INITIAL IMPRESSIONS AND

FINAL ASSESSMENTS OF WRITING QUALITY

by

THOMAS CLIFFORD GARDINER

(Under the Direction of Donald L. Rubin)

ABSTRACT

Holistic scoring of single-sample, timed and unassisted extemporaneous writing remains the prevailing method for large-scale, high-stakes measurements of writing quality. The validity of such holistic scoring has been challenged on grounds that raters read hastily and superficially. Their judgments are thus prematurely determined by their first impressions of student essays, rather than by thorough perusal. Indeed, the process of holistic assessment is of interest as an instance of evaluative reading wherein an overall impression unfolds as readers confront each successive element of text. Three interrelated studies were conducted to investigate judgmental processes of holistic raters as they evaluated individual essays or portfolios of student writing in real time. The studies addressed questions of how early text elements function as contextual or priming mechanisms to affect overall evaluations of essays or portfolios.

Study 1 investigated judgments of raters evaluating essays in the naturalistic setting of a statewide writing competency test. Ratings of experimentally manipulated essays revealed that impressions of quality engendered by high or low error density in the first half of an essay are rarely modified by subsequent changes in writing quality.

Employing computer-assisted data collection, Study 2 investigated unfolding evaluations as raters read essays into which elements of either sophisticated or infelicitous writing had been intruded at specific junctures. The results generally indicated that ratings were higher when infelicities appeared late rather than early in essays.

Portfolios of student writing have been widely proposed as an alternative to single-sample assessments. The results of Study 3 indicated that raters were in fact able to suspend first impressions and rather accurately average the quality of portfolio components into an overall score. The experimental portfolios utilized were, however, more standardized than is typical in portfolio assessments.

In short, raters in large-scale essay assessments were susceptible to their first impressions. In a well-controlled portfolio assessment, however, raters were able to withhold judgment of the whole until they have weighed the quality of each constituent essay.

INDEX WORDS: Writing assessment, Composition quality, Evaluation, Writing portfolios, Raters as readers, Validity of holistic ratings

COMPOSITION EVALUATION AS AN UNFOLDING PROCESS: THE IMPACT OF ESSAY FEATURES ON INITIAL IMPRESSIONS AND FINAL ASSESSMENTS OF WRITING QUALITY

by

THOMAS CLIFFORD GARDINER

AB, Davidson College, 1976

MAT, Vanderbilt University, 1983

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2004

COMPOSITION EVALUATION AS AN UNFOLDING PROCESS: THE IMPACT OF ESSAY FEATURES ON INITIAL IMPRESSIONS AND FINAL ASSESSMENTS OF WRITING QUALITY

by

THOMAS CLIFFORD GARDINER

Major Professor: Donald L. Rubin

Committee: Mark Faust
Sally Hudson-Ross
Patricia McAlexander

Electronic Version Approved:
Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2004

DEDICATION

To my folks, David C. and Mary Elizabeth Neal Gardiner, who taught me long ago to respect teachers, to carry home a stack of schoolbooks on my hip, and to enjoy the neighborhood library up the street.

ACKNOWLEDGEMENTS

I am indebted to my dissertation committee for their patience, encouragement, and guidance in making this dissertation better. Sally Hudson-Ross was gracious enough to emerge from her recent retirement to help one more graduate student cross the finish line. I am especially grateful to her for that kindness and for her close reading of the document and her suggestions for improving it. Mark Faust provided steadfast support of me and this project and was instrumental in helping me to appreciate all points of view regarding institutional writing assessment. Patricia McAlexander, closest to me in teaching and assessment experiences, bolstered my confidence that this research was meaningful and valuable. From the beginning to the end of the project, she also offered good critical questions and suggestions that improved the final product.

It was a privilege to work with Don Rubin, a very smart and very good man. I suppose the basic research question running through this dissertation began with me, but it was Don who had the singular vision to research the question in three integrated studies, the experience to design the studies in such a professional manner, and the patience to guide me, a particularly "gradual" student, through the process. It was a tremendous learning experience to work with Don. I will never forget it.

I am also indebted to my many supportive colleagues at Augusta State University, my professional family with whom I have worked and lived for many years. Foremost among them is Bill Dodd, friend and mentor as well as Associate Vice President for Academic Affairs, without whose encouragement and light touch during the inevitable moments of doubt I could

not have completed this dissertation. I am grateful to him. I also owe a debt of gratitude to a number of other colleagues: Maureen Akins, Associate Director of Instructional Technology, who provided truly invaluable assistance in the planning and execution of the online study; Angela Hodge, who provided critical help with bibliography and document format; Dean Elizabeth House and Department Chair Cindy Craig, who generously provided release time to allow me to conduct research projects and to write; Nabil Ibrahim, who in addition to continual encouragement shared his considerable statistical expertise with me; and Sankar Sethuraman and Dharma Thiruvaiyaru, who graciously helped me with the presentation of statistical data.

I am most appreciative of the assistance of Dr. Kathleen Burk, Assistant Vice Chancellor of Academic Affairs and Director of Regents' Testing Program for the University System of Georgia. Dr. Burk allowed placement of research materials in an actual scoring session of the Regents' Testing Program Essay Test, and in so doing made an enormously important contribution to this dissertation project.

Finally, I must express my gratitude to my family and friends. One realizes as the years go by that it is day-to-day living that is most meaningful in life. Friends and family play the most important role in this regard. I am thankful for the continual support of my brother David and my sisters Edith and Mary Beth. I also appreciate the encouragement of friends Scott Allen, Beth and James Bower, Quentin Davis, Magali Duignan, Cammy Fisher, Bill Gray, Greg Hunter, Margie Mosner, Amy Baltzell Patrick, Bob Reeves, John Sappington, and Jim and Elizabeth Thomson.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND PURPOSE

Issues in Writing Assessment

The conceptual and pragmatic tensions that characterize writing assessment today are attributable primarily to two spheres of influence—(1) educators and (2) testing specialists--that at first aligned alongside each other, and then gradually came to oppose each other, during the first waves of modern writing assessment. (See Yancey, 1999 for a historical account of writing assessment in the U.S.) The evolution of this demarcation is attributable in part to a change in demographics beginning in the early 1950s—an influx of new types of students in schools, students different from those familiar to educators at the time. With this change in demographics came new problems in schools: where to place the students, what to teach them, and how to be sure they learned it. The open admissions policies that swept through public colleges and universities in the 1970s also contributed to the demand for reliable, large-scale assessment in academic areas such as writing. What to that point had been the domain of educators alone became a societal problem of concern to educational administrators and testing specialists. Thus was born a conflict that continues to evolve today. The first three waves of writing assessment--advocating in turn multiple choice tests of grammar, mechanics and usage, then holistic scoring of single writing samples written in standardized conditions, and now portfolio assessment--have over the years merged and overlapped with each other. Yet no resolution has emerged regarding which assessment method should prevail, what the purpose of writing assessment should be, or who is best suited to decide such issues.

Indeed, by now writing assessment has become, in the words of Edward White (1996b), a "site of contention" (p. 301). The prototypical high-stakes assessment of writing for decisions like college admissions or professional credentialing is conducted on the basis of a single essay, written in response to a supplied topic, produced under secure testing situations for. , That prototype has been widely criticized for its failure to reflect writers' authentic capacities (e.g., Hillocks, 2002; Huot, 2002; Moss, 1994b; White, 1998; Wiggins, 1994). Since the 1980s, the growing preference in the literature of educational assessment has been for "alternative," "authentic," "performance," or "performative" assessment, which in the case of writing evaluation has generally meant portfolio assessment (see, for example, Broad, 1994; Camp, 1993; Hillocks, 2002; Huot, 1996, 2002; Moss, 1994b; Tierney, Carter, & Desai, 1991; Wiggins, 1994; Yancey, 1992, 1999).

Most of the dissatisfactions with timed, single-sample, often extemporaneous writing assessments revolve around the disjunction between the product-oriented evaluation of texts produced in standardized testing conditions and writing pedagogy which emphasizes context and process over product (Brossell, 1996; Elbow, 1996; Hillocks, 2002; Huot, 1996, 2002; Yancey, 1992). These dissatisfactions include (1) the contention that holistic scoring procedures sacrifice fuller, more accurate assessment for reliability made possible by raters' submission to scoring guides (Charney, 1984; Moss, 1994a); (2) the argument that these assessments, typically relying on the rapid scoring of one extemporaneous essay, have less validity as measurements of writing ability than assessments of multiple texts produced over time with cycles of feedback and revision (Anson, 1991; Applebee, 1991; Broad, 2003; Camp, 1993; Elbow & Belanoff, 1991; Farr & Beck, 1991; Glaser & Silver, 1994; Huot, 2002; Johnston, 1989; Moss, 1994a; Taylor, 1992; White, 1998); (3) the objection that holistic scoring in particular fails to provide useful

feedback to help students learn and to help teachers teach (Elbow & Belanoff, 1991; Glaser & Silver, 1994; Hillocks, 2002; Huot, 2002; Johnston, 1989); and (4) the argument that such assessments result in writing instructors' "teaching to the test" by having students write short, one-draft extemporaneous essays instead of following a curriculum more in line with current writing theory (Elbow & Belanoff, 1991; Glaser & Silver, 1994; Hillocks, 2002; Huot, 2002; Mabry, 1999; Messick, 1989, 1994; Moss, 1994a). A further objection against these assessments is that they produce failure rates disproportionately skewed against minority students, presumably because of surface errors that may be associated with dialect (Farr, 1996; Kamusikiri, 1996).

The objection to typical writing assessments that most closely motivates the present studies is the contention that single-sample, standardized, rapid holistic assessment procedures miss the intended mark--the accurate measurement of a student's writing ability. If large-scale holistic scoring of texts is to be defended as a valid assessment method, it must answer to the charges that holistic rating procedures constrain raters to a hasty, superficial reading process that causes them to overvalue surface correctness, undervalue such features as voice and creativity, and miss textual subtleties that warrant closer, more careful consideration before reaching an assessment decision (see, for example, Charney, 1984).

Given the volume of such criticisms against this variety of writing assessment, it might be expected that assessment practices in colleges and universities would have undergone substantial change. However, Sandra Murphy's 1993 survey (discussed in Elbow, 1996 and White, 1996c) of postsecondary writing assessment practices, commissioned by the Conference on College Composition and Communication (CCC), revealed that portfolios were being used in only three percent of large-scale writing assessments in colleges and universities, while

assessments using timed, standardized writing samples accounted for forty-nine per cent of these assessments. Typically, such assessments are used for purposes of placing students into what appears to be the most appropriate level of composition courses or of certifying the competency of students to exit out of a program or a certain level of instruction. For example, at a typical large college or university, readers may annually process many thousands of essays to place or exit students or to decide their competency as writers. Furthermore, as of the late 1990s, a nearly equal number of so-called assessments of writing ability at colleges and universities did not even examine actual student writing, but used indirect evaluation methods to ground decisions for placement or certification purposes; 48 per cent of such assessments were based on multiple-choice tests of students' knowledge of grammar, mechanics, and usage (Elbow, 1996; White, 1996c). Thus, the prevailing practice in post-secondary, institutional, large-scale writing assessment has not gravitated powerfully to portfolios, despite the shift in preference reflected in the professional literature.

To be sure, it is uncertain what direction large-scale writing assessment has taken in the decade since the CCC survey. Nor is it clear what direction writing assessment may take in the future. In fact, Brian Huot (2002) has called for a "transformation of writing assessment identity" (p.14), in which portfolio practitioners in schools create site-based and locally controlled writing assessments sensitive to curricular context and independent of statistical validation associated with traditional writing assessments. Huot cites validation procedures employed by William L. Smith (1993) at the University of Pittsburgh and Richard Haswell (2001) at Washington State University as models for grounding such local assessments.

Nevertheless, to this point, writing assessment remains the site of contention described by White, with accountability-based institutional goals still in conflict with the educational goals of

writing teachers who advocate assessments sensitive to conditions that promote effective teaching and learning of writing. To understand why large-scale institutional assessment practice has not yet shifted more strongly toward portfolio assessment, it is necessary to examine in turn the reservations about authentic writing assessment which have emerged in the literature: (1) Reservations about achieving agreement among raters in scoring portfolios (Hamp-Lyons and Condon, 1993; Freedman, 1993; Herman, Gearhart, & Baker, 1993; Messick, 1994; Myers & Pearson, 1996; Nystrand, Dowling, & Cohen, 1993; White, 1998) persist. Note, however that some portfolio advocates (Huot, 2002; Johnston, 1989; Moss, 1994a, 1994b) argue that scoring reliability is not a necessary condition for validity in the assessment of writing ability. (2) Conducting authentic assessment is logistically complex and financially expensive due to the time-consuming nature of the process, compared to the greater efficiency of standardized assessments (Madaus & O'Dwyer, 1999; Williamson, 1993; White, 1998). (3) Consumers of writing assessment data cannot be sure whose work is being assessed in portfolios which include collaborative pieces or essays written without supervision (Clark, 1993; Hamp-Lyons & Condon, 1993; Herman, Gearhart, & Baker, 1993; Underwood & Murphy, 1999; White, 1998). (4) Moreover, to this point, research on summative portfolio assessment has failed to substantiate claims of the effectiveness of portfolios for distinguishing levels of writing proficiency (Freedman, 1993; Larson, 1996; White, 1998). (5) Referee or judge bias and inconsistency across tasks and standards prevent comparability among portfolio assessments (Myers & Pearson, 1996).

The field of writing assessment is thus in the uncomfortable position in which the prevailing theoretical trend in writing assessment has appeared to shift to portfolio assessment, but institutional practice in large-scale contexts outside the classroom has not followed (Huot,

2002). A likely explanation for this disparity is that conflict among the various stakeholders has produced a politically charged atmosphere and impeded a widespread shift to portfolio evaluation in high-stakes, large-scale assessment. (For discussions about the influence of politics on policy and practice in writing assessment, see White, Lutz, and Kamusikiri, 1996; Hillocks, 2002; Huot, 2002.) As White (1996a, 1996b, and 1996c) has observed, writing teachers are by no means the only or necessarily the most powerful stakeholders involved in establishing assessment practices. They, and others with a stake in writing assessment, may be uneasy about large-scale holistic assessment, but that method continues to be used on a widespread basis (Williamson, 1993 and White, 1996). Thus, there is a need to investigate more carefully the nature of the holistic assessment process to shed further light on challenges to the validity of this often discredited but still prevailing practice. Although composition research has burgeoned in the last three decades, questions about the validity of holistic assessment constitute one area in which there remain important gaps in empirical research.

The process of large-scale holistic rating poses some inherent problems. Holistic raters (1) operate under the duress of reading large numbers of papers in short periods of time and (2) are constrained to adhere to a carefully designed set of scoring guidelines. Yet few empirical studies have examined holistic rating as a process that occurs in real time, with sequential patterns of behavior, both conscious and unconscious. Indeed, in his review of the literature on direct writing assessment, Huot (1990) pointed out the need for more investigations into the nature of evaluative reading processes if we are to better understand issues of validity in writing assessment. For instance, although holistic raters are theoretically presumed to base their judgments of writing quality on the prescribed criteria they have been trained to use, their judgments may be distorted by fatigue or by the order in which they read papers (Braddock,

Lloyd-Jones & Schoer, 1963; Barker, Kibler, & Hunter, 1968; Hamp-Lyons & Condon, 1993). Further, a small number of studies (Harris, 1977; Rafoth & Rubin, 1984) have produced evidence that raters' attention can be subconsciously pulled away from prescribed evaluative criteria by mechanical errors in writing, a powerful influence on raters' perceptions of overall text quality and even distorting readers' perceptions of other textual features.

When one considers essay rating to be one specialized form of reading, it appears quite likely that order effects and effects of perceptually salient text features like mechanical errors might exert undue and/or irrelevant impact on perceptions of composition quality. Skilled readers bring to these texts expectations about characteristic discourse structures that influence their interpretations of those texts. They have strong predispositions to search new texts for discourse structures which frame content. Frank Smith (1994) theorized that skilled readers progressively formulate and reformulate interpretations of a text's meaning as they read, basing their interpretations on genre schemes and text structures characteristic of similar texts. Some knowledge-based comprehension theorists (see, e.g., Frederiksen, 1986; Ruddell & Unrau, 1994) have suggested that a reader's expectations about text structure may have a top-down, expectation-driven effect on the reading process as textual schemata influence meaning-making. Flower (1988) theorized that skilled readers actively read to construct a framing purpose when they read essays and base their reading of the rest of the essay on this initial interpretive framework. Such readers, she observed, experience discomfort and struggle to create meaning when they can't find appropriate cues in a text to help them construct a satisfactory framing purpose.

Extrapolating from that view of essay rating as skilled reading, it seems plausible to suppose that when holistic raters are asked to assess a large number of student compositions

rapidly, the ease or difficulty of comprehending text structures might figure prominently in the sequential formulation of raters' evaluative impressions, and formation of final judgments of writing quality, as they read each composition. Empirical investigations of how frequently, and at what point in the process of reading a composition, raters encounter difficulty in comprehending a writer's message might yield insights into the influence of text structures on ratings of writing quality. In general, there is a need to study, in real time, the impact of various textual features, including text structures, on judgment formation as readers score texts holistically.

The general interest in investigating the dynamic, sequential nature of holistic assessment should logically be extended to portfolio assessment as well. Too little is known about the ways in which portfolio raters arrive at judgments of writing quality. As Richard Larson (1996) has observed, the literature on writing portfolios consists largely of declarations of preference, reports of teachers' experiences with them, and recommendations for their use. Just as there is value in conducting research into questions of validity in holistic assessment of texts produced in standardized conditions, so is there the need to investigate empirically whether various models of portfolio assessment of writing deserve claims made for their validity as measures of writing ability (Larson, 1996; White, 1998). Since holistic scoring is the most commonly used rating method in portfolio assessment, it follows that researchers should design studies of this method of portfolio assessment.

An important question regarding the validity of portfolio assessment for college placement and certification decisions is whether the judgments of raters asked to assess portfolios holistically will be variably influenced by the presence in the portfolio of a writing sample composed under standardized conditions. Compositions written in standardized

conditions are sometimes included in portfolios as a check of the authenticity of other pieces produced by students over longer periods of time and with the benefits of feedback and revision (Clark, 1993, and Herman, Gearhart, & Baker, 1993). Although there have been very few empirical studies of holistic assessment of portfolios, a small handful of studies have produced evidence that holistic raters of portfolios do not score the portfolios systematically. Hamp-Lyons and Condon (1993) found that raters did not attend systematically to every text in the portfolio; some essays exerted greater salience than others, some were largely ignored. Herman, Gearhart, and Baker (1993) found that overall holistic ratings assigned to portfolios did not match the averages of scores the raters had given to pieces comprising the portfolio. Nystrand, Cohen, and Dowling (1993) were able to achieve acceptable levels of inter-rater reliability among holistic raters, but only when individual texts comprising each portfolio were rated separately. Nystrand and colleagues also concluded that single holistic portfolio scores are valid only when portfolios have been standardized in certain respects. This small body of empirical investigations of portfolio assessment processes needs to be expanded.

Purpose

This dissertation consists of a coordinated set of three studies collectively investigating questions which bear upon the issue of construct validity of holistic assessments of both single essays and portfolios. The general focus of the studies was the nature of the holistic reading process. More specifically, the studies investigated the real-time dynamics, both conscious and subconscious, of the reading processes of holistic raters as they assess student texts across large-scale, high-stakes assessment contexts.

The type of writing assessment with which these studies were concerned was large-scale, summative assessment; the scholastic level of the assessment was post-secondary. The specific

purpose of the assessments was to assure a minimally acceptable level of competency in writing ability established by the Georgia state university system. Student performance on the assessment, which is required of all baccalaureate degree-seeking students in this university system, carries high-stakes consequences. Receiving a passing score on the assessment clears the student to receive a degree when required course work is completed; failure on the assessment usually requires a student to enroll in a writing course, often a remedial, non-credit course, and to retake the writing assessment at the conclusion of that course. Students who fail are required to continue this cycle until they pass the writing assessment. In these respects, this writing assessment program functions in a fashion quite similar to other system-wide or institution-wide programs commonly installed to police at least one aspect of general education requirements expected by accreditation bodies and by the community at large.

Thus this form of assessment carries tremendously high stakes for students: validation and a sense of accomplishment for those who pass, disappointment and duress for those who do not. These consequences are significant and should not be taken lightly. In addition to these ramifications, however, there is the institutional goal of assuring quality of instruction and learning. If the assessment functions properly, it helps ensure appropriate levels of each throughout a statewide university system, a circumstance which is presumably good for students as well as other stakeholders in this publicly supported enterprise. In light of both of these perspectives, it is incumbent upon writing professionals and assessment professionals to examine closely questions of validity regarding such high-stakes assessments. The goal should be to ensure that such assessments are functioning properly, and if they are not, to take appropriate steps to remedy the situation.

The three projects described in this dissertation were designed to contribute to the body of research focusing on such questions of validity. Considered as a package, the three projects investigate, in greater breadth and depth than is possible in individual studies, questions about the behaviors of raters engaged in holistic assessment of writing quality. The common objective of the studies was to examine closely the tendencies of raters engaged in the specialized type of impressionistic reading employed in holistic writing assessments.

Two of the studies investigated the extent to which impressions formed in holistic raters' minds by selected textual features manipulated at strategic points in a single text are associated with their final judgments about the whole text. Study 1 investigated the rating behaviors of readers evaluating essays in the naturalistic context of a high-stakes writing competency test in the actual statewide university system assessment. Study 2 investigated the unfolding evaluations of raters as they read essays from start to finish, in real time, in a laboratory context. Study 3 extended the investigation of how raters' judgments develop sequentially to a portfolio assessment context; it investigated the question whether evaluative impressions formed by exposure to extemporaneous essays, placed at varying locations in the portfolios, influence raters' final judgments about the entire portfolios.

Research Questions

The following are the research questions investigated in the studies:

STUDY 1

1.0 To what extent are raters' judgments of an entire essay in a large-scale testing context influenced by the placement of errors (early versus late) in the essay?

STUDY 2

2.0 To what extent are raters' sequential, paragraph-by-paragraph evaluative impressions of an essay in a laboratory assessment context influenced by the location of infelicities (early or late) in the essay?

2.1 To what extent are raters' sequential, paragraph-by-paragraph evaluative impressions of an essay in a laboratory assessment context influenced by the rhetorical sophistication of the introductory paragraph of the essay?

2.2 To what extent are raters' sequential, paragraph-by-paragraph evaluative impressions of an essay in a laboratory assessment context influenced by the presence or location of organizational consistency (consistency, inconsistency at paragraph two, inconsistency at paragraph four) in the body of an essay relative to the organizational plan established in the first paragraph of that essay?

2.3 To what extent are raters' sequential, paragraph-by-paragraph evaluative impressions of an essay in a laboratory assessment context influenced by the density (early versus late) of error in the essay?

STUDY 3

3.0 Given a portfolio consisting of three writing samples, does the inclusion of a writing sample of higher or lower quality than the other two samples affect the ratings of the portfolio as a whole?

3.1 Does the position of the dissonant writing sample in relation to the other two samples influence the rating of the portfolio as a whole? That is, does either the primacy or recency of a dissonant quality writing sample within a portfolio affect the overall evaluation of that portfolio?

Definitions of Terms

Assessment. Defined broadly, "the term *assessment* ... includes the idea of evaluation, the use of assessment information to judge or appraise the knowledge or skills of those who are the subject of the assessment" (White, 1996b, p. 1). In this dissertation, assessment will refer to summative assessment unless the term is otherwise qualified.

Large-scale assessment refers to formal assessment procedures beyond the individual classroom and beyond the classes of an individual teacher, either summative or formative, which seeks to judge or appraise the ability of individual students or groups of students for an institutional purpose, whether the scope is local, statewide, or national. When the ability or skill of individual students is the focus of large-scale assessment, the purpose may be to decide placement or exit or to measure achievement. When the target focus of the assessment is the performance of a group of students, the purpose may be the evaluation of a program or institution.

Holistic. Borrowing from Charles R. Cooper's overview (1977) of holistic evaluation of writing, the term holistic is used in this dissertation to refer to any evaluation of written texts in which each text or portfolio of texts is ranked or sorted impressionistically according to a set of guidelines established for a particular assessment. The evaluation may take the form of scoring, grading, or categorizing for placement or for measurement of proficiency. However, the term holistic is restricted to evaluative procedures which require a rater to judge the quality of a text or collection of texts as an overall entity. Holistic assessment requires the rater to balance in his or her mind the different criteria specified in the grading rubric while evaluating a text or portfolio and to assign only one score, rank, or placement category to that text or portfolio.

Portfolio. The term portfolio will be used broadly here to refer to any collection of written texts representing the work of a writing student. A portfolio may include multiple drafts of one or more documents, final drafts (only) of different compositions, or a combination of the two approaches. The writing tasks and texts may be selected by the student (the preference of many portfolio advocates) or specified by a teacher or other institutional official. Further, writing portfolios may (and usually do) include compositions written over time after cycles of feedback and revision and compositions written under standardized conditions. While many writing portfolios include an essay in which the writer reflects on process and product, the portfolios described in Study 3 of this dissertation did not contain such a text. In Study 3, described in Chapter 5, all portfolios used consisted of one essay purportedly composed under standardized conditions and two essays purportedly written in freshman composition courses with cycles of feedback and revision.

The term **standardized conditions** as used here denotes controlled conditions in a writing assessment context. Writing assessments conducted in standardized conditions typically hold uniform such variables as time allowed for the writing task, topic(s) available to the student, and reference materials allowed. Furthermore, standardized conditions ensure that the text produced by the writer is the work solely of that writer.

Introductory organizational frame. The term *introductory organizational frame* refers here to the overall organizational plan for the essay as it is presented explicitly in the first paragraph of the essay. When an organizational frame is made explicit in an introductory paragraph, it consists, minimally, of a thesis statement which presents the basic, controlling idea for the rest of the essay. An introductory paragraph's organizational

frame may also include, either in the thesis or in subsequent sentences, a concise presentation of the major supporting ideas to be developed in the body paragraphs of the essay.

Rhetorical sophistication of the introductory paragraph refers to the strategy employed by the writer in the introductory paragraph to engage the reader's interest and to present the thesis and organizational plan of the essay. There are references in Chapter 4 to several successful college writing textbooks which present rhetorical explanations of how writers can structure introductory paragraphs to take advantage of the psychology of reader interest. Chapter 4 explains how rhetorical sophistication of the introductory paragraph is operationalized in the studies.

Consistency/Inconsistency between introductory organizational frame and subsequent essay paragraphs refers to the consonance or dissonance between an organizational frame of the main ideas supporting the essay's thesis, predicted in an essay's introductory paragraph, and the actual organizational structure of ideas presented in the subsequent body paragraphs of that essay.

Limitations of the Studies

For all studies, generalizability of the writing samples used as the basis of measurements of writing ability is open to question. Whether of the expository or persuasive genre, whether single-sitting writing samples written on demand or longer samples composed over time, the writing samples included in these studies might be challenged as to whether they are representative of other samples belonging to other genres or different writing conditions. Further, the generalizability of raters' behaviors in the studies might be questioned; whether other raters, even those with similar training and experience, would behave similarly is open to

question. In defense of the methods used here, on the other hand, all essays and raters were associated in some respect with a genuine state-wide postsecondary writing assessment.

In addition, the motivation of raters in Studies 2 and 3 must be considered as a limitation. Because these studies are laboratory studies, the psychology of raters involved can not be the same as if they were actually engaged in authentic, high-stakes assessments. The settings in which writing evaluations were conducted posed a limitation in Studies 2 and 3, which were laboratory studies (in one case, administered on computer). Study 1, in contrast, strove for greater ecological validity; experimental essays were interspersed among genuine rating tasks in an actual statewide assessment.

CHAPTER 2

REVIEW OF THE LITERATURE

Issues of Reliability and Validity in Large-Scale Holistic Assessment

One can hardly expect to be measuring what one claims if two observers cannot even agree that they are seeing the same thing. Much of the literature on writing assessment during the 1960s and `70s was in this respect preoccupied with reliability in assessments of writing quality. Diederich, French, & Carlton (1961), who initiated research on holistic essay rating on behalf of the Educational Testing Service, stated that any conclusions about writing ability drawn from essay grades were almost certain to be unwarranted unless agreement among raters was first established. Difficulties in developing reliable procedures for direct assessment of writing led many institutions to rely instead on objective (typically multiple-choice tests) to measure "writing ability."

Objections to indirect measurement of writing ability (e.g., Braddock, Lloyd-Jones & Schoer, 1963), led many in the field of writing assessment to develop standard scoring practices for the scoring of essays in direct writing evaluation. Godschalk, Swineford and Coffman (1966) found evidence that acceptable reliability and validity were possible when multiple writing samples and multiple readings (scorings) were used in assessment. Diederich (1974) developed analytic scoring procedures based on the construction of a scoring rubric that enabled researchers and practitioners to achieve acceptable levels of inter-rater reliability. Lloyd-Jones (1977) advanced both theory and valid practice in direct writing assessment, calling for evaluators to

develop a separate scoring guide suited to the rhetorical traits being evaluated in the context of each assessment.

Veal and Hudson (1983) summarized the following observations about the reliability of the three dominant direct writing assessment procedures of the 1970s. Analytic scoring had proved to be the most reliable of all direct writing assessment procedures. Primary-trait scoring procedures achieved somewhat lower reliability correlations, and the extra time required made this method more costly. (Nevertheless, the National Assessment of Educational Progress still uses versions of primary-trait scoring in large-scale writing assessment; see Hillocks, 2002 and White, 1998.) Holistic scoring achieved acceptable reliability levels and correlations with analytic scores. Issues of reliability, then, had been an important, perhaps preoccupying, concern of writing assessment research and practice for several years by the early 1980s. The historical context for this overriding concern with scoring reliability in the 1970s and 80s was a direct outgrowth of the "frustrating combat with the perfect scoring reliability of the machine-scored tests" of multiple-choice, indirect writing assessments (White, 1993, p. 84).

Questions about the construct validity of these three direct assessment methods, however, became prevalent in the 1980s. Gere (1980) suggested that these evaluation methods, in restricting raters' attention to selected and targeted features of texts, actually "subsumed" the natural reading/evaluation procedures of raters. Charney (1984) echoed the objection that holistic scoring procedures distort normal evaluative reading processes and suggested that these procedures might cause assessors to sacrifice valid judgments for interrater reliability.

Another challenge to validity arises because raters of a student text often attempt to assess the student's writing ability rather than the actual quality of the text itself, introducing excessive variability in judgments across raters (Martin, 1987; Barritt, Stock, & Clark, 1986).

Moreover, abundant evidence suggests that rater expectations and biases about evaluating writing sometimes interferes with applying the rating rubric evaluators were supposed to be following (Barritt, Stock, & Clark, 1986; Breland & Jones, 1984; Rafoth & Rubin, 1984; Scannell & Marshall, 1966).

The questioning of the validity of commonly employed direct writing assessment methods intensified in the 1990s. Calls for alternative assessments of writing have centered on the concept of "authenticity." Proponents of authentic assessment point out that assessments of single writing samples written under pressure in standardized test conditions do not even approximate the ecological conditions of writing in the world beyond "standardized" test rooms (Hillocks, 2002; Huot, 1990, 1996, 2002; Lucas, 1992; Mabry, 1999). Many teachers of writing, joined by an increasing number of authorities in the field of educational testing and measurement (e.g., Elbow & Belanoff, 1991; Huot, 2002; Moss, 1994a,1994b; Taylor,1994; Wiggins, 1994; Yancey, 1992) have called for authentic performance assessment. The hallmark of authentic writing assessment allows for (1) multiple writing tasks and genres, (2) extended time (including cycles of feedback and revision), (3) student selection of writing topics and tasks, and (4) evaluation by the teachers involved in the actual delivery of instruction, teachers who know the students personally.

Williamson (1993) describes the current state of the field of writing assessment as "tearing itself loose" from a psychometric foundation and establishing an assessment foundation based in writing theory. Bob Broad has declared that “the age of the rubric has passed” (2003, p. 4). Huot (2002) has openly called for a transformation of writing assessment grounded not in psychometric procedures “that attempt to fix objectively a student’s ability to write…based on an outdated theory supported by an irrelevant epistemology” but rather in emergent new ideas about

measurement and validity more in keeping with ecological theories of writing pedagogy and curriculum (p. 94).

Validity of writing assessment, then, does not mean the same thing to writing professionals as it did a decade or two ago. One of the leading voices in the new conceptualization of assessment, Pamela Moss (1994), proposed the development of alternative assessment methods without reliability across raters or writing tasks to supplement existing assessment procedures. She acknowledged the findings of Nystrand, Cohen, and Dowling (1993) that some portfolio assessment procedures have difficulty achieving either type of reliability. She also acknowledged that having teachers involved in their own students' performance assessment contributes to construct-irrelevant variance in measurement (see Resnick & Resnick, 1992), and thus undermines the interests of some educational stakeholders. Nevertheless, Moss argued that psychometrically reliable assessment methods silence the voices of those (teachers) who are most knowledgeable about the context of the learners' situation and of those (students) who are most affected by the assessment. Moss (1994) contended that a disciplined, collaborative systematic evaluation of multiple writing performances that "honors the lived experiences" of teachers and students can achieve a meaningful validity as long as the procedure invites challenges to initial interpretations and an insistence on supporting evidence for any revised interpretations.

Moss (1994) and others advocating a broad, nonpsychometric conception of validity based their proposals on what Messick (1989) and others have called "consequential validity" and Camp (1993) has referred to as "systemic validity." The notion of consequential validity assumes that assessment inevitably drives curriculum and pedagogy. The problem, therefore, with psychometrically reliable writing assessment methods is that they lead to invalid teaching

and learning activities, such as the production of one-sample, one-draft student compositions. Moreover, according to proponents of consequential validity, psychometrically based assessment methods have not protected some groups (e.g., different dialect speakers) whose backgrounds put them at a disadvantage when fair opportunities for demonstrating the general construct of "good writing" are not adequately offered.

Messick (1994) concurred that issues of consequential validity should have a place in discussions regarding appropriate choices of methods in writing assessment. But he diverged from Moss's (1994) position by maintaining that psychometric reliability does matter in all assessment procedures that claim generalizability. Reliability across tasks is important, according to Messick (1994), because the meaning of the target construct is tied to the range of tasks represented in the assessment. Hence, in this view, proponents of alternative assessments such as portfolios need to address such circumstances as the inability to establish score reliability across writing tasks (Nystrand, Cohen, and Dowling, 1993) or to replicate student proficiency levels in subsequent assessments.

Furthermore, Messick (1994) contended, empirical evidence that “authentic assessments” actually do ensure construct validity must be presented. In this view, face validity is not sufficient. Messick maintained that there has been insufficient empirical evidence to support the claim that complex performance tasks such as lengthy, collaborative writing tasks across different genres can adequately represent the construct of individual writing ability. In at least some few studies, portfolio assessment systems have in fact demonstrated acceptable levels of interrater agreement (e.g., Underwood & Murphy, 1998). Nonetheless, the overall record of portfolio assessment used in large-scale contexts has frequently been marked by problems, including excessive time and cost burdens as well as weak reliability (Ramirez, 1999).

Holistic Raters' Responses to Various Textual Features

One key issue regarding the validity of composition rating—whether traditional or "authentic"—pertains to the salience of specific writing features in affecting raters' judgments of overall composition quality. This is an issue of validity because it might be the case that a particular textual feature—the occurrence of surface error, for example, or use of contractions in formal essays—so overwhelms raters' judgments that those raters are unable to attend to other criteria laid out in a rating rubric. In such a case were found to be true, then ratings would reflect not overall composition quality, but rather the presence of those particular textual features. The perceptual dominance of textual features is not an unlikely supposition. After all, it is much easier for a reader—particularly a reader in the position of making holistic judgments about dozens of essays at a sitting—to focus on concrete text features like misspellings than on more amorphous qualities like development and tone.

Following the trail of raters' reactions to textual features and gauging the impact of those reactions on their judgments of overall text quality was one of the dominant concerns of composition research during the 1970s and 80s. Mina Shaughnessy (1977) pointed out the almost infinite variations of evaluative response to student texts: the rapid shifting of interior reader response to the innumerable permutations of strengths and weaknesses arranged in different sequences.

Most of the research done on the influence of textual features on judgments of writing quality has been conducted using holistic scoring of short essays written in a single sitting. Within this assessment context, Freedman and Calfee (1983) found that although other factors such as topic and rater training do affect raters' judgments (for better or for worse), the most

significant influence on the skilled rater is the text itself. But which textual features exert the most influence on these judgments has proven to be a rather complex question.

Some issues in this area of research have proved less muddied than others. Composition length (usually operationalized by word count) is well established as a predictor of rated composition quality (Breland & Jones, 1984; Nold & Freedman, 1977; Grobe, 1981; Stewart & Grobe, 1979). Perceived neatness and attractiveness have also correlated robustly with judgments of writing quality (Markham, 1976; Breland & Jones, 1984). Lexical diversity and lexical choice (e.g., the density of Latin-derived words as opposed to Germanic-derived) likewise predict holistic ratings (Grobe, 1981; Nielsen & Piche, 1981; Nold & Freedman, 1977; Stotsky, 1979, 1981). Even an isolated appearance of an epithet in an essay can trigger perceptions of a writers' gender (sometimes inaccurately), and thereby also affect perceptions of writing quality (Haswell & Haswell, 1996).

The interest in measures of syntactic complexity in the 70s and early 80s stimulated several investigations to see whether this feature of writing was associated with higher quality ratings. These studies did not produce conclusive evidence of such a relationship, however, at least not across writers' developmental stages (Huot, 1990). Several studies with younger (elementary-age) students found a positive relation between syntactic maturity (as measured by T-unit length) and writing quality (Veal, 1974; Stewart & Grobe, 1979; Witte, Daly, & Cherry, 1986), though this association may only at the low end of the syntactic maturity scale. Many other studies, those involving high school and college students' writing, produced no evidence that syntactic complexity is associated with judgments of overall writing quality (Nold & Freedman, 1977; Nielsen & Piche, 1981; Crowhurst, 1980; Stewart & Grobe, 1979).

Similarly, the nature of the relation between writing quality and textual cohesion/coherence that emerges from empirical research is unclear. While some studies found a positive association between density of cohesive ties and holistic judgments of composition quality (Witte & Faigley, 1981; Neuner, 1987), at least one other (McCulley, 1985) found no such linkage.

Amount of content and degree of organization have frequently been linked together by researchers investigating the influence of these closely related textual elements and writing quality. Several researchers have found evidence that content and organization figure more prominently in raters' judgments about the quality of texts than do any other textual features such as mechanical errors (Freedman, 1979; Freedman & Calfee, 1983; Breland & Jones, 1984; Pula & Huot, 1993).

However, the way in which researchers operationalize content is a sometimes overlooked variable when comparing studies and drawing conclusions about the influence of these text elements on quality judgments (Rafoth and Rubin, 1984). Amount of content has often been operationalized simply as length (e.g., Freedman & Calfee, 1983), but of course other facets of development are suggested by the term "content:" quality of ideas, sufficiency of support, sufficiency of detail, and so on. Rafoth and Rubin (1984) operationalized semantic load in texts by manipulating the density of propositions. To determine the relative impact of proposition density, they also manipulated another textual feature, mechanics. Their ensuing results contradicted previous findings that content is a more powerful influence than matters of mechanical correctness on evaluative judgments about writing quality. Even when raters were explicitly instructed to disregard writing mechanical errors in evaluating essays, their judgments of quality were significantly influenced by this textual feature. The results of their study

supported similar findings by Scannell and Marshall (1966). Harris (1977) also found that even though raters expressed a preference for weighting content and organization more heavily than error, their judgments of quality were significantly influenced by mechanical errors.

This contradiction warranted further discussion of the relative salience of textual features--the variable influence, at a subconscious level, of one or more features of a text on raters' judgments of overall quality in that text. Breland and Jones (1984) found discrepancies between holistic raters' perceptions of the textual features that influenced them and the actual influence of those features. Rafoth and Rubin described this effect as a kind of diffusion or contamination: a rater's perceptions about the quality of one textual feature or even the quality of the overall text are subconsciously distorted by reaction to a different textual feature. They questioned whether raters are capable of constraining themselves to adhere to assigned scoring rubrics.

This construct of relative salience has implications for studying the impact on quality ratings of other textual features besides error and content density, features such as organizational patterns. In a study of the habits of skilled readers as they read essays, Linda Flower (1988) found that they actively constructed a framing purpose for each text and read with this initial interpretive framework through the rest of the essay; inversely, they struggled to make meaning when they did not find a discernible framing purpose. It is reasonable, then, to hypothesize that raters of student texts in large-scale assessment contexts search for organizational cues as they evaluate these texts. A strong association has been observed between the use of structured organizational patterns and higher quality ratings when raters evaluate argumentative texts (Connor, 1987 and Ferris, 1994). Personal experience expositions which include embedded

narratives received higher ratings than texts which were written in response to the same topic, but which were organized as pure narratives (Hake, 1986).

These relatively few investigations of the influence of organizational features of texts on rater behavior in large-scale assessment contexts leave questions that need to be explored. If readers naturally search for a discernible framing purpose as they read, then is it possible that holistic raters form strong impressions about texts on the basis of organizational cues, or the lack thereof, in early portions of those texts? What if any implications are there for raters' judgments about texts which delay presentation of controlling idea until late in the text or texts which lack explicit presentation of controlling idea? And what are the implications for rating papers which lack congruence between the organizational plan presented in a first paragraph and the subsequent organization of ideas in the rest of the composition? Further, might a rater's reaction to the organizational features of a text be so strong as to have greater salience for overall judgments of writing quality than is called for by the rubric within a particular assessment context?

Evaluative Reading Processes

One area where reading research and writing research intersect is in the examination of reading processes involved in writing assessment. Although it is difficult to establish a clear historical link between reader-response theory and theoretical trends in writing assessment, there are some undeniable similarities in focus. The attention which Louise Rosenblatt (1938, 1978) fixed on the pivotal role of the reader in constructing literary texts is mirrored in some writing assessment theorists' focus on the crucial role of the reader evaluating student writing. Raymond (1982), for instance, makes this theoretical link explicit in saying that it is the marriage of text and reader that produces a judgment about the quality of the writing. While much of this

attention to reader response in writing assessment circles has focused on formative evaluation of student writing -- its effects on student motivation, revision practices, and improvement -- some work on holistic assessment of writing shows evidence of a reader-response orientation.

One reader-response construct that seems to have informed theory in holistic writing assessment is Stanley Fish's concept of the interpretive community (1980a; see also discussion in White, 1998), which suggests that an individual's perceptions of and judgments about a text are shaped by the socially constructed and assimilated assumptions shared by the group to which one belongs. White contends that Fish's construct is what allows us to integrate reader-response theory into large-scale writing assessment. The concept of the interpretive community, he suggests, "rescues" from a dangerous subjectivism the theory that individual readers create meaning as they proceed through a text (p.99). As has been indicated previously, however, while some theorists have affirmed the ability of those engaged in various assessment contexts to evaluate student writing using a shared set of static criteria (Freedman & Calfee, 1983; Huot, 1993; Lloyd-Jones, 1977; Martin, 1987), others have questioned the capacity of raters to fully internalize grading rubrics in ad hoc assessment contexts and to base their assessment decisions purely on those prescribed criteria (Barritt, Stock, & Clark, 1986; Breland & Jones, 1984; Harris, 1977; Rafoth & Rubin, 1984). For this latter group, the danger of individual rater bias looms large.

Another promising avenue of research indicated by reader-response theory is suggested by Stanley Fish's detailed reconstruction of the moment-to-moment process of reading and interpreting a literary text (1980b). His close attention to the developing responses of the reader proceeding line by line through a poem suggests that the crucial question in interpreting a text is not what the text means, but how readers make meaning. Fish's emphasis on the individual

variability of meaning-making in reading the belles lettres is by no means paralleled in the kind of specialized reading involved in writing assessments that place a premium on scoring reliability. However, his close attention to the psychology of the reader suggests the need for investigations into the real-time nature of the holistic reading process, especially given the fact that many assessments have high-stakes consequences for students. There has been a scarcity of research in which holistic rating is examined as a sequential process that occurs in real time. The psychology of holistic readers--including the various conscious and subconscious processes at work as they proceed through different types of texts, in various assessment contexts--is a rich area of inquiry for the future. Indeed, Brian Huot has issued a call for much more extensive investigation of the nature of the reading process of those involved in writing assessment (1990).

In the literature of writing assessment, the best known model of the evaluative reading process of holistic raters has been proposed by Freedman and Calfee (1983). They constructed an information-processing model of composition rating in which they identified three separable sub-processes that underlie the evaluative reading of a composition:

1. Raters read to comprehend text and build a "text image." Every reader builds a slightly different text image, but for a homogeneous group of raters, they suggest, the similarities should outweigh the differences.

2. As the reader builds a text image in working memory, s/he simultaneously evaluates the text image and begins storing impressions into long-term memory. These impressions are the beginnings of evaluative judgments. Freedman and Calfee call these interior, unspoken impressions “covert evaluative judgments.” They claim that their studies suggest that a skilled, trained group of evaluators store similar text images and share common values about the texts.

3. Once the reading of the text has been completed, the rater articulates the evaluation, usually with a letter grade or numerical score. Whether a substantive rationale or explanation of the score is offered (e.g., end-paper comments to students) depends on the purpose of the evaluation, for instance, whether it is formative or summative.

As for the relation among sub-processes, Freedman and Calfee (1983) theorized that they normally occur in an ordered sequence but allowed that the process may be recursive. Current reading theory, they noted, favors the recursive model. To account for the specific focus of a reader's attention at any given moment, they posited a monitoring entity which directs attention to appropriate concerns during reading; they did not, however, go into detail about this latter facet of the evaluative reading process.

Aside from the features of the individual text itself, the Freedman and Calfee (1983) model accounts for two types of variables influencing the formation of raters' judgments: (1) rating task environment variables, e.g., time, length, physical environment, training for the specific assessment task, the purpose of the assessment, and the audience for the assessment; and (2) personal rater characteristics, e.g., reading ability, world knowledge, values about writing and about the subject matter of topics, and expectations about writers, constraints on writers in the assessment context, and texts produced in that context.

The adaptability of the Freedman and Calfee (1983) model to subsequent theory and research is assured in large part because it acknowledges the influence of nontextual variables on raters' judgments. Freedman and Calfee preceded other theorists (e.g., Flower, 1994; White, 1994) who similarly maintained that raters' expectations about writers and constraints of the writing assessment context influence (1) raters' willingness (or unwillingness) to make inferences about the writer's intentions beyond the surface of the text and (2) the evaluative

impressions formed as they read. Their model also anticipated research into the background factors which influence raters' bedrock attitudes and convictions about writing that they bring to ad hoc writing assessment contexts (see, e.g., Huot, 1993; Pula and Huot, 1993; and Martin, 1987).

Freedman and Calfee (1983) concluded that under proper assessment conditions, the text being assessed exerts the greatest influence on the skilled evaluator. In an assessment context in which personal rater characteristics are controlled, presumably by training in the application of the particular scoring rubric being used, the skilled evaluator stores text images that primarily reflect variations in texts rather than in rating task environment variables or rater characteristics.

There are, of course, limitations to the Freedman and Calfee (1983) model. It presumes that readers can consciously restrict their attention to prescribed rating criteria as their judgments are forming. As noted previously, this presumption has been challenged (Barritt, Stock, and Clark, 1986; Breland and Jones, 1984; Harris, 1977; Rafoth and Rubin, 1984). The model does not address the possibility that individual rater characteristics simply cannot be controlled out of the process of judgment formation. The theory of relative salience, for example, which suggests that rater judgments may be subconsciously distorted by differential perceptions of textual features (Rafoth and Rubin, 1984) had not yet been advanced by the publication of the Freedman and Calfee model. And though their model does account for the influence of task environment variables and rater characteristics on ratings, it does not explicitly hypothesize how those factors interact to create the dynamic aspect of the holistic rating process. For example, it does not specifically describe the possibilities that raters' judgments might be distorted by fatigue or by the order in which papers are read, two of many concerns which might be addressed by research which investigates holistic rating as a process that occurs in real time.

Holistic Rating as a Reading Process That Occurs in Real Time

A number of research questions might be fruitfully explored with methodologies that facilitate real-time, or at least sequential, investigations of holistic scoring procedures. Because large-scale assessment involves intense reading and concentration by raters over prolonged periods of time, there is the question whether rating decisions are distorted by fatigue. Also relevant to the question of distortion is the sequential nature of the holistic scoring process. The varying order in which texts are read may produce undesirable effects on rating decisions. For example, individual raters may compare one text against others recently read and arrive at a different judgment than they would have if the order had been different. Furthermore, different raters, reading papers in differing orders, may rate some texts lower or higher than their colleagues because of these unintended order effects. This phenomenon has been confirmed empirically in the context of evaluating speeches as they are delivered (Barker, Kibler, and Hunter, 1968) and in the context of portfolio assessment (Hamp-Lyons & Condon, 1993.

A slim body of empirical research focuses on the nature of the holistic rating process as it occurs over time. These studies have mainly employed methodologies of think-aloud protocol analysis, text annotations, and interviews to investigate these questions. Two studies investigated issues pertinent to the criticism that holistic rating interferes with the natural reading processes of raters by restricting their attention to those criteria contained in the assessment rubric. The concern these studies addressed is that. the natural, personal, and perhaps more valid response of a reader is sacrificed for the sake of reliability (Charney, 1984). Huot (1993) and Pula and Huot (1993) found evidence that contradicted that theory. Whereas novice raters did devise *ad hoc* evaluative strategies as they rated texts, veteran holistic raters not only read more efficiently, they appeared to have organized their training and professional background

experiences into more coherent rating strategies than novice raters. Further, protocol analyses in Huot's study revealed that veteran raters were much more likely than their counterparts to reserve judgment about the quality of texts until after finishing reading each text. In both studies, researchers concluded that results lent support to the contention that holistic rating procedures, appropriately executed, can sustain validity in writing assessment.

Wolfe (1997) also used protocol analysis to study the rating behaviors of proficient and nonproficient holistic scorers. Like Huot and Pula and Huot, he found that raters able to apply a scoring rubric with high levels of interrater agreement tended to suspend judgment of writing quality until later in the essay. Less proficient raters made evaluative decisions earlier in essays and more frequently during the course of their reading. Wolfe speculated that less proficient raters expended more energy and broke up the evaluation process while reading because they were not well prepared to use the scoring rubric.

Other studies, in contrast, have produced evidence that holistic rating procedures do not necessarily ensure valid assessment. Martin (1987) used protocol analysis, text annotation, and interviews to conduct investigations into the process of reading student texts to make course placement decisions about entering college students. She concluded that holistic raters making placement decisions on the basis of these texts read subjectively, in quite different fashion from one another. The findings of this study indicated that these raters read not to evaluate the quality of each text, but to imaginatively construct the writer behind the text in an attempt to judge the best course placement. Martin suggested that even when raters arrived at the same placement decisions, it was questionable that they could claim validity in their assessments of writing ability. In a very similar vein and using similar methods, Vaughan (1991) found evidence that holistic raters did not internalize scoring guidelines that they applied uniformly to each text they

evaluate. Despite similar training, the raters in this study focused on different textual features as they read, some (e.g., handwriting) not even included in prescribed guidelines.

This small body of research leaves a number of important concerns about the nature and integrity of the holistic reading process underexplored in real-time or sequential frames. At what point in the course of reading an essay, do raters' evaluative impressions assume status as tentative or final judgments about the quality of a text? Are there other textual features besides mechanics which occupy greater prominence in raters' perceptions as they proceed through texts and register impressions about writing quality? What are they?

One theoretical thread pertinent to this discussion is Frank Smith's (1994) description of the comprehension processes of skilled readers. Although Smith has greatly expanded the scope of his reading theory in other treatments (1997, 2003) to include cultural allusion, competence, and other pedagogical concerns, in Understanding Reading he theorizes that skilled readers, as they read a new text, progressively create meaning by making a series of predictions about the text based on their familiarity with conventional and characteristic genre schemes and discourse structures of similar texts. These internalized cognitive structures, according to Smith, provide readers with frameworks for interpreting new texts. Reading, then, is a process of trial and error, of predictions made and then modified or discarded and replaced with more predictions, as the reader encounters subsequent portions of the text. The more successfully writers enable readers to anticipate formal structures, Smith theorized, the more successful the reader's comprehension. Inversely, the more discrepant the interaction between a reader's predictions and the formal structures of a text, the less successful the reader's understanding of the text. We can extend this principle to composition assessment contexts, including high-stakes, large-scale assessments of impromptu student essays. Smith's model suggests that texts which provide (or fail to provide)

structural cues facilitate (or undermine) in raters' minds not only comprehension, but also unfolding evaluations of writing quality.

Freedman and Calfee's (1983) analysis of reading for assessment in particular, and Smith's (1994) general model of expert reading as a "psycholinguistic guessing game," seem particularly well suited to describing the processes of workaday raters plying their craft in situ. Far from ideal readers for whom real-time constraints and self-presentation issues are moot, raters engaged in typical placement examination or competency testing contexts must read in ways that are often more expedient than ideal. Because of these constraints on large-scale raters reading evaluatively, Huot (1993) has specifically urged more studies of raters operating in realistic rather than in laboratory settings.

Concerns about External Validity in Writing Assessment Research

Methodological criticisms of research on writing assessment raise questions about the external or ecological validity of findings. Observing that much of this research was conducted in non-naturalistic laboratory contexts, Huot (1993) has pointed out that this research may be challenged as lacking the validity of studies done in naturalistic assessment conditions. Because the real-life consequences of rating decisions in high-stakes contexts are missing in laboratory studies, it is possible that raters' mental processes and judgments about texts are altered. Thus, it becomes suspect to generalize about the nature of large-scale, high-stakes assessments from the results of laboratory studies. To preserve the high-stakes ecology of large-scale assessments, quasi-experimental research designs should be employed to explore the reading processes of holistic raters in actual field assessments. For instance, distortion of raters' judgments due to the phenomenon of relative salience in their perceptions of textual features might be investigated by

manipulating selected features of a sample of essays and then submitting these essays to naturalistic field assessment procedures.

Furthermore, many studies of holistic assessment have required raters to rate relatively few texts, thereby failing to duplicate large-scale assessment conditions and control for the possibility that fatigue affects rating decisions. In addition, there has been a scarcity of studies that have investigated the possibility of order effects in the sequencing of essays to be rated, although many field practitioners and some theorists (Braddock, Lloyd-Jones, & Schoer, 1963) have assumed that this variable does influence rating decisions. Many critics of large-scale assessment have charged that fatigue distorts the decision-making processes of raters.

Studies of rater behavior have often relied on protocol analysis methodology as a window into cognitive processes. However, these studies can be challenged on the grounds that their findings may be compromised by the self-awareness of raters reporting their mental processes in a research context. While protocol analysis is useful for investigating rater behavior in real time contexts, this methodology cannot provide a transparent window into the formation of raters' judgments because of the artificiality this methodology introduces into these normally private processes (See Smagorinsky, 1989, 1998) for a discussion of problems in using protocol analysis in writing research).

Accordingly, there is a need to conduct research which addresses these methodological concerns. To overcome the intrusiveness of think-aloud data collection, it is desirable to explore methodologies that permit raters to register judgments in real time, without interrupting their judgmental processes. This naturalistic research could be coupled with comparative laboratory studies designed to trace sequential impression formation in real time as raters evaluate these same manipulated essays. One such method is to present raters with texts in controlled fashion

on a computer screen, having them report ratings as they register in their consciousness (by means of a mouse) as they read in real time. The foci of these laboratory studies should be (1) whether the independent variable textual features distort raters' judgments and (2) whether varied placements of these textual features within texts exert different degrees of influence on rater judgments of overall writing quality.

Empirical Research on Portfolio Assessment

A further criticism of traditional large-scale holistic assessment of writing is that it usually has involved the evaluation of only one writing sample, typically an extemporaneous essay written under standardized conditions in a high-stakes test context. This type of assessment, say many critics, is not reflective of what we know about how writing is produced in naturalistic conditions, with opportunities for collaboration and revision over time. Thus, many argue, such traditional assessments lack validity as fair assessments of writing ability (Applebee, 1995; Elbow and Belanoff, 1991; Farr and Beck, 1991; Glaser and Silver, 1994; Moss, 1994a, 1994b).

Another objection to such traditional assessment practices is based on the tension created when such assessment tasks retroactively influence curricular and instructional practices. This phenomenon, for years referred to as "teaching to the test" or "washback" from testing to curriculum (Rubin & Mead, 1984), is now drawing interest from measurement specialists who describe the issue as a matter of "systemic" or "consequential" validity, maintaining that the validity of an assessment method depends in part on the impact the assessment procedures have on instruction (Camp, 1993; Messick, 1989, 1994; Moss, 1994a, 1994b).

Partially in response to these critiques, a number of theorists and practitioners have called for the development of "authentic" assessment procedures more attuned to the processes of writing than are assessments of texts conducted in standardized conditions. Portfolio assessment

has been the most frequently proposed alternative assessment method (Huot, 2002; Ramirez, 1999; White, 1998; Yancey, 1999). Its value within the individual classroom, especially for purposes of formative assessment, is well accepted. However, the ongoing debate over the validity of portfolios for formal assessments makes it incumbent on empirical researchers to continue investigating portfolio assessment if it is to be defended as a method of evaluating writing for high-stakes, summative purposes (Larson, 1996).

In addition to difficulties of establishing reliability between raters and reliability across writing tasks in portfolio assessment discussed earlier in this chapter, the assessment literature identifies several other issues relevant to questions of portfolio validity. The assumption that portfolios, providing as they do the opportunity for writers to collaborate and revise over time, do a better job of measuring the writing process than standardized assessments has been questioned (Hamp-Lyons & Condon, 1993, Larson, 1996). Portfolios that include only final drafts instead of multiple drafts may privilege writers in one way (i.e., their missteps are not held against them), but do not evaluate writing skill over the process of writing. Another question regarding the validity of portfolio assessment is the issue of whose work is being assessed when the portfolio includes collaborative compositions – the student's or the collaborator's? (For discussions of the issue of authorial authenticity in portfolio assessment, see Clark, 1993; Elbow & Belanoff, 1991; Herman, Gearhart, & Baker, 1993; Wolcott & Legg, 1998.)

The most significant question about portfolio assessment in the context of this study has to do with uncertainty regarding the nature of the evaluative process in assessing portfolios, i.e., how assessors arrive at holistic ratings of portfolios. Hamp-Lyons and Condon (1993) found that holistic raters did not systematically attend to every text in a portfolio as they evaluated. Instead, holistic raters frequently reported arriving at a decision on a portfolio's score during their reading

of the first or second of the four texts included in the portfolio. They observed that "readers seemed to go through a process of seeking a 'center of gravity' and then read for confirmation or contradiction of that sense" (p. 182). This sampling or order effect was so pronounced that students were advised to organize their portfolio texts in order of quality, placing first the text of highest quality and then arranging the others in order of descending quality. Herman, Gearhart, and Baker (1993) found that holistic raters' overall scores of portfolio quality exceeded the averages of scores they gave to individual texts included in each portfolio. Further, texts produced in standardized (supervised and timed) conditions were consistently scored lower than texts written over time and with feedback. The holistic scores assigned to the entire portfolio consistently exceeded the ratings assigned to the standardized texts. The researchers were unable to determine whether these results indicated that standard writing performances underestimated students' writing ability or whether other factors, such as help received by students on the nonstandardized written pieces, accounted for these disparities.

Investigating Holistic Portfolio Assessment as a Dynamic Evaluation Process

One approach to deciding which investigations should be undertaken about the process of rating portfolios holistically in large-scale, high-stakes contexts is to consider questions that have emerged from the work already done on traditional holistic evaluation. In applying these questions to portfolio assessment, of course, one essential difference between traditional holistic scoring contexts and portfolio assessment contexts must be taken into account: the difference in the number of texts being read in each situation. When just one rating is assigned to a collection of texts, what implications are there for research into the process of portfolio assessment?

One crucial question which has emerged from empirical studies of traditional holistic assessment is whether raters adhere to the grading criteria established within a particular

assessment context. The work of Hamp-Lyons and Condon (1993) and Herman, Gearhart, and Baker (1993) suggest that some holistic raters of portfolios do not direct their attention evenly to each of the texts included in a portfolio. Theoretically, the phenomenon of relative salience suggested by Rafoth and Rubin (1984) might account for this inconsistency. But there may be different ways in which a reader's perceptual inconsistency might operate over the process of reading several texts before deciding on one holistic judgment. Does a rater's powerful evaluative reaction to one text, to a cluster of texts, or to a textual feature (or cluster of features) in one or more texts diffuse to other texts in the portfolio and exert undue influence on the rater's evaluative judgment of the whole portfolio? Does a rater register different evaluative reactions for the different texts in a portfolio and then balance and even-handedly consolidate those different impressions according to specified guidelines into a global judgment of writing quality? Clearly, empirical research needs to be done on the process by which raters formulate judgments about portfolios in real time.

Certainly, the possibility that order effects may affect evaluative judgments during portfolio assessment merits investigation. Just as it is possible that impressions formed while reading the early portions of a single text might influence a rater's impressions of subsequent portions of that text, so might a reader's evaluation of one text in a portfolio influence his or her evaluations of a subsequent text (or texts), or even the entire portfolio itself.

These gaps in knowledge about portfolio assessment could have important implications in high-stakes assessments. For instance, some portfolio assessments follow the practice of including in the portfolio one extemporaneous essay written under standardized conditions as a gauge of the student's independent writing ability (e.g., see Clark, 1993, and Herman, Gearhart, & Baker, 1993). It is reasonable to expect that the placement of that extemporaneous essay --

first in the portfolio, for instance, or last -- could influence the rater's overall judgment of the quality of writing in the portfolio.

One specific area worthy of investigation into holistic portfolio assessment, then, is the question of the impact which extemporaneous essays placed at different locations within portfolios have on holistic ratings of overall portfolio quality. Such research, like other investigations proposed earlier in this chapter, would investigate holistic assessment, in this case portfolio assessment, as a dynamic process that should be studied as a real-time evaluative process.

Directions for New Research

Describing the fluid interior responses of those engaged in reading to evaluate student texts is difficult. The permutations of clusters of textual features alone are innumerable; considering additional variables such as topic, evaluation context, and rater differences makes the prospect of delineating the sequential nature of evaluative reading processes a daunting one indeed. Nevertheless, there may be, in the sequential reading processes of raters involved in various assessment contexts, general tendencies of impression formation which bear on the important question of validity for each of those assessment contexts.

Although Freedman and Calfee's (1983) descriptive model of the evaluative reading process provides a useful base for considering new questions about rating as a real-time phenomenon, there are gaps in the Freedman and Calfee model. Their description of the activity of the reader's executive monitor presumes conscious control in mediating between the sub-processes of building text image and formulating evaluative judgments. Empirical studies done subsequent to the publication of Freedman and Calfee's model, however, suggest that holistic raters are not always conscious of their responses to texts they evaluate. Neither does the Freedman and

Calfee model address the question whether there are sequential patterns of response, either conscious or unconscious, from beginning to end of the evaluative reading process. Because some empirical studies suggest that unintended patterns of response do occur in holistic rating, the validity of this assessment method is called into question. Hence, there is a need for further empirical investigations of the moment-by-moment process of rating texts holistically.

Frank Smith's (1994) characterization of the comprehension processes of skilled readers as a sequential meaning-making process also has implications for new empirical investigations of the evaluative reading process. Smith's theory suggests that readers create meaning progressively by making a sequence of predictions about the text based on familiar genre schemes and discourse structures which provide readers with frameworks for interpreting new texts. Reading, then, is a process of trial and error, of predictions made and then modified or discarded and replaced with more predictions, as the reader encounters subsequent portions of the text. We can extend this theoretical description of reading for comprehension to the evaluative reading processes of raters in writing assessment contexts. It is useful to consider the premise that raters of student compositions approach the assessment task with internalized frameworks for evaluating those compositions. It is also useful to investigate the theory that raters make predictions about the quality of student texts as they read them, and then confirm or modify those impressions as they encounter subsequent portions of text. This characterization has potentially important implications for raters in large-scale, high-stakes assessment contexts, who are under duress to evaluate a high number of texts in a short time span.

It is quite natural to expect that a portrait of actual holistic rating processes would depict raters as functioning at least in part stochastically. That is, veteran raters operate under the twin pressures to (1) maximize efficiency (in terms of both speed and accuracy) and (2) maintain

concentration and minimize boredom (by seeking stimulation and giving short shrift to the formulaic). As a result, they may sample only portions of a writing sample rather than reading it carefully from beginning to end. More specifically, raters may form an initial probability assessment of overall composition quality based on early parts of the writing sample--the first two or three paragraphs, perhaps. They may then sample selectively from the remainder of the text in order to confirm or disconfirm this early probability assessment.

Current research has not adequately addressed questions about the validity of holistic reading processes in large-scale, high-stakes writing assessments. There is a relatively scant body of empirical research to investigate the criticism that such raters read hastily and superficially before they reach an assessment decision. There have been only a few studies which have specifically addressed the question whether holistic raters form early judgments about texts and then read superficially to confirm or disconfirm those judgments. To be sure, Huot (1993) concluded that trained holistic raters reserved judgment about the quality of a text until after reading the entire composition; however, this study did not address two important questions: first, whether raters formulate private, tentative judgments about writing quality as they read silently and independently in many large-scale assessment settings; and second, if they do reach preliminary judgments, the extent to which these tentative judgments influence final judgments.

The set of studies comprising this dissertation was devised to help address these emerging areas of inquiry in composition assessment.

CHAPTER 3

STUDY 1

This chapter describes the participants, the test instruments used, the methods for collecting the data, the procedures for analyzing the data, and the analysis and results for Study 1. This study investigated the rating behaviors of readers evaluating essays holistically in the naturalistic context of a large-scale, high-stakes test of writing competency in an actual statewide, university system writing assessment. It was designed to answer criticisms that studies of composition rating in laboratory settings cannot elucidate composition rating that takes place under the pressures and stresses of authentic large-scale assessment conditions (Huot, 1990). In this study, the most obvious and well-established predictor of composition quality ratings--that is, error density (Harris, 1977; Rafoth & Rubin, 1984; Scannell & Marshall, 1966)—was manipulated. To ascertain the potential impact of initial impressions versus the potentially modifying effects of later-appearing textual features, two versions of essays were developed. In one version, a high density of error preceded low error density, and in the other version a relatively low error rate was followed by a high density of errors.

The following research question was investigated in Study 1:

RQ1.0: To what extent are raters' judgments of an entire essay in a large-scale testing context influenced by the placement of errors (early versus late) in the essay?

Method

Study 1 investigated the rating behaviors of readers evaluating essays holistically in the naturalistic context of a large-scale, high-stakes test of writing competency in an actual

statewide, university system writing assessment. To test for effects of early-induced expectations versus late-stage information on raters' judgments about the overall quality of an essay, it was desirable to construct essays that differed in quality in specifiable ways from first half to second half. In this study, half the essays were characterized by high quality in the first half of the essay and then low quality in the second half. The other half of the essays were characterized by low quality first, then high quality in the latter half of the essays. Because of the documented effect of error rate on judgments of quality, quality was operationalized here in terms of error density.

Participants

The 88 participants in this study were drawn from the pool of raters employed by the University System of Georgia in the Regents' Testing Program Essay Test (RTPET) in the summer of 1998. Participants were drawn randomly from the six pools of raters who gathered at the six scoring sites situated across the state during the summer 1998 RTPET scoring sessions. These raters were all employed as composition instructors among the 34 universities, four-year colleges, and two-year colleges that make up the University System of Georgia. This population of raters, operating within the context of the RTPET, was chosen in order to make as naturalistic as possible the conditions and rating procedures used in assessing the essays manipulated for the study.

The Regents' Testing Program trains participating raters at each formal RTPET session to adhere to prescribed holistic scoring guidelines in order to promote acceptable levels of inter-rater reliability (to examine these RTPET scoring guidelines, see Appendix A). In addition, the Director of Testing for the University System of Georgia annually advises all participating raters of their percentage of agreement with other raters for RTPET rating sessions and stipulates that acceptable levels of agreement (typically 88% or better agreement with at least one of the two

other raters of each essay) must be achieved for individual raters to continue to participate in these assessments. All participants in this study had established levels of interrater reliability that met or exceeded the standards for the RTPET.

Participating raters were drawn randomly from the six different pools of raters who gathered at the six state scoring sites used in a summer 1998 RTPET scoring session. Three different raters assessed each of 40 experimental essays. Normal RTPET procedures assured that no rater would score the same essay more than once.

Instrumentation

The Regents' Testing Program Essay Test. The RTPET is a sixty-minute test of student writing ability taken by all bachelor's degree-seeking students in the University System of Georgia (For documentation about the Regents' Test, visit the web site: http://www.gsu.edu/~wwwrtp/). Students write one essay extemporaneously on one of four topics presented to them at the beginning of the sixty-minute writing period. The test is designed to ensure an acceptable level of competency in writing ability as established by the University System of Georgia.

RTPET essays are scored by three raters using a four-point holistic scale. A rating of 1 on this scale signifies a failing essay; a score of 2 signifies a marginal pass; a score of 3 signifies a higher level of competency; and a score of 4 signifies a still higher level of competency. For a student to pass the RTPET, that student's essay must receive scores of 2 or higher on this scale from at least two of the three readers who rate it.

Stimulus Essays. The forty experimental essays that were included in the study were constructed from twenty original student essays written for a previous RTPET administration. The experimental essays were constructed so as to control for comparability in holistically rated

overall quality. Since the critical decision in the RTPET is whether an essay passes or fails, all of the source essays used in constructing the experimental essays were selected on the basis of their previously rated proximity to the 1 (failing) to 2 (marginally passing) range. Five of the source essays had originally received scores of 1, 1, 1 from three raters using the Regents' Program Essay Test holistic scale; five essays had received ratings of 1, 1, 2; five essays had received ratings of 1, 2, 2; and five essays had received ratings of 2, 2, 2.

These original essays were manipulated to control for location of density of error. This independent variable had two levels, early versus late density of error. As a preliminary step to constructing experimental versions of essays, frequency counts of errors and error types were performed for two samples of RTPET essays. Since it would be unrealistic for either half of an RTPET essay to be entirely free of errors, a background level of error, "low" error rate, was established. This low error rate was established by sampling error frequencies representative of marginally passing essays. To sample these error frequencies, ten essays were selected that had been rated 2, 2, 2 in a previous RTPET session and frequency counts were conducted. To establish a "high" error rate, frequency counts of error were conducted on ten essays which had received failing ratings (1, 1, 1) from all three raters in the same RTPET administration. All frequency counts were conducted by two researchers working independently on the two sets of ten sample essays; coding differences were resolved through collaborative review and agreement on appropriate error coding. Means of error count per essay (rounded to the nearest whole number) were calculated for each type of error counted in each of the two samples. The means for each type of error are presented in Table 3.1.

Table 3.1

Representative Number of Errors in Sample of Failing and Marginally Passing RTP Essays

Type of Error	Frequency, Failing Essays	Frequency, Marginally Passing Essays
Misspelling	6	2
Fragment	1	0
Faulty Parallelism	1	0
Subject-Verb Agreement	2	0
Verb Tense Shift	1	0
Uninflected Verb (-d,-ed,-ing)	2	0
Unidiomatic Phrasing	3	0
Wrong Word Usage	1	0
Word Missing	1	0
Wrong Form of Word (a/an, possessive error, plural form error, adjective/adverb)	5	1
Missing Comma	4	1
Misused Comma	2	0
Misused Semicolon	1	0
Shift in Person	1	0

Using the frequency counts of the two samples of essays and the error hierarchies of Hairston (1981) and Leonard and Gilsdorf (1990), a selected group of errors representative of marginally passing essays and another group of failing essays were selected to be systematically instantiated in the experimental essays used in the study. In order to accentuate the contrast between the errors representative of the marginally passing essays and those errors representative of the failing essays, the number of errors selected for each group were adjusted: the total

number of errors selected to be representative of low error rate were reduced in number, while some of the higher-gravity errors selected as representative of high error rate were increased in number. The two groups of errors judged to be representative of low rate and high rate of error are presented in Table 3.2.

Table 3.2

Manipulated Error Distribution Constituting High and Low Error Density in Stimulus Essays

Error	Frequency, High Rate of Error Density	Frequency, Low Rate of Error Density
Faulty Parallelism	1	0
Sentence Fragment	1	0
Subject-Verb Disagreement	2	0
Verb Tense Shift	1	0
Wrong Word Usage	1	0
Wrong Form of Word	2	1
Missing Comma	3	1
Misused Comma	1	0
Shift in Person	1	0
Misspelling	3	1

Two experimental versions of each original essay were constructed. One version concentrated the errors representative of failing essays in the first half of the essay, while the errors representative of marginally passing essays were dispersed in the second half of the essay. The second experimental version of each essay inverted this pattern: low error rate in the first half of the essay, high error rate in the second half. Word counts were used to determine the midpoint of each essay. Errors were dispersed throughout each half of every essay according to the high and low density scheme. The dispersal of errors was nonsystematic; however, the editors relied on their experiences as veteran teachers of writing to instantiate and disperse these

errors in realistic fashion. All 20 pairs of experimental essays controlled placement of error in similar fashion.

Controlling errors in this way, in a sample of essays that in original form received ratings in the failing or marginally passing range, allowed examination of the potential effect of early or late error placement on the two outcome variables: ratings of writing quality on the RTPET four-point scale and the RTPET pass-fail decision for each essay.

Procedures

To preserve the illusion for RTPET raters that the experimental essays were authentic student essays, each was copied by hand on official testing paper provided by the Regents' Testing Program. Five individuals copied eight essays each — two experimental versions of each of four source essays. To equalize differences in penmanship across levels of writing quality, each copyist prepared handwritten versions across all four levels of quality in the original source essays. Thus, each of the five individuals copied two experimental versions of one essay that received initial ratings of 1, 1, 1; two experimental versions of an essay that received initial ratings of 1, 1, 2; two versions of an essay initially rated 1, 2, 2; and two versions of an essay initially rated 2, 2, 2.

During the summer quarter of 1998, with the cooperation of the state-wide RTPET testing coordinator, the forty experimental essays were non-systematically dispersed in stacks of authentic RTPET student essays and distributed among the six regional sites across the state where RTPET rating sessions were conducted. In order to keep assessment conditions naturalistic, neither site coordinators nor raters were notified that experimental essays were included among the essays to be rated. However, essays were distributed to the six RTPET scoring sites across the state in such a manner that no raters were exposed to both experimental

versions (weak first half, strong second half; strong first half, weak second half) of any experimental essay. In accordance with routine RTPET procedures, each experimental essay was scored by three raters using the Regents' Test four-point holistic scale.

Analysis

The outcome measures used in the statistical analyses of Study 1 were the three ratings (1-4) for each experimental essay and the pass-fail decision for each essay. A passing essay, conforming to authentic RTPET practice, was one that received at least two out of three ratings of 2 or higher. A failing essay was one that received fewer than 2 scores of two or higher. The research question lent itself to analysis of variance procedures. A single 3 (rater) x 2 (error placement) ANOVA with repeated measures on both rater and error placement was conducted for the RTPET ratings.

In addition, a 2 (error placement) x 2 (pass/fail) contingency table was constructed and tested for significant association by means of the Chi-square statistic.

Results

The raw data appear in Table 3.3. Although no inferential test of statistical significance was applied to these raw data, a descriptive comparison of essays characterized by high error rate in the first half against essays in which high error rate was delayed until the second half reveals that essays with high density of error in the first half received more failing ratings (21.7% vs. 16.7%) and far fewer merit ratings (8.3% vs. 28.3%) than did essays characterized by low error rate in the first half.

Table 3.3

Raw Ratings of RTP Essays with Manipulated Error Densities

Location of Error Density	Frequency of "1" Rating	Frequency of "2" Rating	Frequency of "3" Rating	Frequency of "4" Rating	Total
High Density, First Half	13	42	5	0	60
High Density, Second Half	10	33	15	2	60

A summary of the ANOVA appears in Table 3.4. It indicates a significant main effect for rater ($F_{2,38} = 4.16$, $p < .05$) as well as for location of error density ($F_{1,38} = 5.03$, $p <. 05$) . While the rater variable was of no theoretic interest, it does indicate 18% of variance in RTP ratings attributable to rater differences. Location of error density, the variable of interest here, accounted for 21% of the variance among RTP scores. The mean rating for the papers with high error density preceding low error density ($M = 1.87$) was significantly lower than for the papers with low error density preceding high error density ($M = 2.15$). This finding is consistent with a conclusion that raters were most influenced by errors appearing in the first half of the essays, and that their impressions, once formed, were not readily altered.

Table 3.4

Summary of ANOVA

Source	Sum of Squares	Df	Mean Square	F	p-value	Partial Eta^2
Rater	1.617	2	.808	4.160	.023	.180
Location of Error Density	2.408	1	2.408	5.033	.037	.209
Rater * Error	.317	2	.158	.563	.574	.029
Error (Rater * Error)	10.683	38	.281			

The 2 x 2 cross-tabulation of pass decision with error density location appears in Table 3.5. As it indicates, each of the two error conditions contained just four failing essays. The associated Chi-square statistic could not be reliably calculated, since the expected value of half

the cells was less than the rule of thumb $n = 5$. Nonetheless, it is clear that no association between error location and pass rate existed in these data.

Table 3.5

Cross-tabulation, Location of Error Density and Pass/Fail Decision

Pass/Fail Decision	Location of High Error Density		
	Early in Essay	Late in Essay	Total
Failing Essays	4	4	8
Passing Essays	16	16	32
Total	20	20	40

Preliminary Discussion of Study 1

The most important finding in Study 1 is the significant difference in mean ratings assigned to essays with high error rate in the early part of the essay as opposed to essays with the same types and frequencies of error, but with those errors delayed until the latter part of the essay. This finding suggests that raters operating under conditions of duress typical of large-scale, high-stakes assessment rating sessions were more likely to be influenced by language written in the first half of essays than what was written in the second half.

Although the data in this study do indicate a propensity for error location to affect impressions of overall composition quality, that propensity was not dramatic enough to impact high-stakes pass/fail decisions.

Thus the data provide some support for the hypothesis that early impression formation may exert a disproportionate influence on holistic ratings of essay quality when raters are functioning in large-scale assessment contexts. It is important to note, however, that this study manipulated texts to be different only in terms of error, a textual feature that has previously been demonstrated to exert fairly powerful influence on holistic ratings of writing quality. It remains to be seen whether raters would respond similarly to essays manipulated differentially vis-à-vis early text versus later text.

CHAPTER 4

STUDY 2

This chapter describes the participants, the test instruments used, the methods for collecting the data, the procedures for analyzing the data, and the analysis and results for Study 2. Study 2 manipulated selected textual features—location of error density, occurrence and/or location of problem in essay organization, and sophistication of introductory strategy--in order to further test the impact of initial impressions on evolving and final judgments about the whole text. Unlike Study 1, which examined the rating behaviors of readers in an actual statewide university system assessment, Study 2 investigated the unfolding evaluations of raters from paragraph to paragraph as they read these texts from start to finish, in real time, in a laboratory context.

The following are the research questions which were investigated in Study 2:

RQ 2.0 To what extent are raters' sequential, paragraph-by-paragraph evaluative impressions of an essay in a laboratory assessment context influenced by the location of infelicities in the essay (early or late)?

RQ 2.1 To what extent are raters' sequential, paragraph-by-paragraph evaluative impressions of an essay in a laboratory assessment context influenced by the rhetorical sophistication of the introductory paragraph of the essay?

RQ 2.2 To what extent are raters' sequential, paragraph-by-paragraph evaluative impressions of an essay in a laboratory assessment context influenced by the presence or location of organizational consistency (consistency, inconsistency at paragraph two,

inconsistency at paragraph four) in the body of an essay relative to the organizational plan established in the first paragraph of that essay?

RQ 2.3 To what extent are raters' sequential, paragraph-by-paragraph evaluative impressions of an essay in a laboratory assessment context influenced by the density (early versus late) of error in the essay?

Method

Study 2 investigated the rating behaviors of readers evaluating essays holistically in a laboratory setting. To test for effects of early-induced expectations versus late-stage information on raters' judgments about the quality of an essay, experimental essays were constructed in which specific textual features were manipulated at strategic points so as to trace raters' developing impressions as they encountered those features. In this study, three textual features –introductory strategy, error density, and organizational consistency– were manipulated. Introductory strategy was operationalized in terms of sophistication of rhetorical strategy; error density was operationalized in much the same manner as was employed in Study 1; and organizational consistency was operationalized in terms of adherence to, or departure from, the organizational plan presented at the end of the introductory paragraph.

The study used a computerized, online presentation of essays one paragraph at a time, with raters required to report a tentative judgment of writing quality before seeing each new paragraph within an essay. This method ensured sequential collection of raters' impressions of writing quality in an approximation of real time.

Participants and Instrumentation

In this study, 288 composition judgments were rendered by 12 raters, each of whom was an experienced teacher of freshman composition at his or her college or university within the

University System of Georgia. Further, all were experienced raters of essays written for the Regents' Testing Program Essay Test (RTPET), a 60-minute test of extemporaneous student writing ability taken by all degree-seeking students in the University System of Georgia. All study participants had participated in numerous state-conducted training sessions to enable them to evaluate student essays written for the Georgia RTPET with the Regents' Testing Program holistic scoring guidelines. All participants had established interrater reliability that met or exceeded the standards established for the RTPET (R. Keithley, personal communication, April 29, 2000). Participants were financially compensated for their involvement in the study with a sum identical to the daily remuneration paid to RTPET raters.

The RTPET is designed to ensure an acceptable level of competency in writing ability taken by all degree-seeking students in the University System of Georgia. A rating of 1 on this scale indicates a substandard or failing performance; a score of 2 signifies a minimal passing performance; a score of 3 signifies a clearly passing performance; and a score of 4 indicates superior performance. (For an explanation of the holistic scale used in the Regents' Testing Program Essay Test, see Appendix A.)

Stimulus Essays

A total of 42 essays were used in the study, including 24 experimental essays and 18 filler essays. The experimental essays were constructed by modifying a set of 24 essays written on RTPET topics. Some of these original essays used as sources for the experimental essays were selected from a pool of authentic RTPET essays written by university system students for a previous administration of the RTPET. Other essays were either written by university students under standardized conditions similar to those used in the RTPET and then modified by the

researchers, or in a few cases in which essays with the needed parameters could not be located, were written by the researchers themselves to simulate student essays.

The experimental essays were sampled or created so as to control for comparability in a number of respects. Each of the 24 original essays used as sources for the experimental essays had received ratings of 2, 2, 2 (minimal pass) on the RTPET holistic scale from all three raters who scored them. Those that were written in a previous administration of the RTPET were scored by raters operating within the normal conditions of that test administration. The others were submitted to three veteran Regents' Test raters, each of whom rated the essays independently. The rationale for selecting original essays all rated at the minimal passing level was to measure the potential of the experimental manipulations for changing the pass-fail status of the essays. There were additional steps to control for comparability. To avoid differences in handwriting, all essays (experimental and filler) were word processed. Each experimental essay consisted of five paragraphs, including one introductory paragraph, three body paragraphs, and one concluding paragraph. A common feature of the introductory paragraphs used in the study was the inclusion of a tripartite framework of ideas at the end of the opening paragraph suggesting the organizational framework of the rest of the essay. And since essay length has been shown to have a powerful influence on ratings (Breland & Jones, 1984; Nold & Freedman, 1977; Grobe, 1981; Stewart & Grobe, 1979), all essays selected fell within a 20 per cent range from the mean number of words (M = 443, range = 370-554).

Experimental essays were manipulated to vary three independent variables: degree of sophistication of rhetorical strategy (high versus low) in the opening paragraph, degree of consistency between organizational cues in the opening paragraph and organizational cues in subsequent paragraphs (consistency, inconsistency at paragraph two, or inconsistency at

paragraph four), and placement of error (early versus late placement of error). Each of these versions reflected one of the 12 treatment combinations created by the different levels of the three independent variables (2 x 3 x 2) of interest (that is, 2 levels of error density location by three levels of organizational violation by 2 levels of introduction sophistication). Among the 24 experimental essays, each of the 12 treatment combinations was represented twice (the replication factor), so that it would be possible to generalize any significant effects beyond the idiosyncratic impact of a particular essay.

The 18 filler essays were included to make the rating task more realistic. The filler essays broadened the range of essay quality beyond the experimental essays, which clustered around the minimally passing rating benchmark. Each filler essay had been predetermined to reflect one of three quality ratings surrounding the minimal passing performance rating of 2 on the four-point holistic scale used in the Regents' Testing Program Essay Test. Ratings had been assigned to the 18 filler essays prior to their inclusion in the study. Those which were actual RTPET essays had been rated in the naturalistic scoring procedure used in the Regents' Testing Program. Other filler essays were submitted to veteran RTPET raters, and the consensus quality ratings of those essays were determined by these raters (via agreement of the two raters initially consulted, or two out of three in cases where the first two raters had reached a split decision and a third rater was consulted). To function as distractors from the experimental essays, which were constructed from essays rated as minimal passing performances, the filler essays were selected because their quality ratings fell either above or below the minimal passing rating. Six filler essays had been predetermined to be failing essays, each having received a consensus quality rating of 1; six had been assigned a consensus rating of 3 (clearly passing); and six had received a consensus rating of 4, signifying a superior performance. The purpose of including these filler

essays was to ensure that the set of essays rated by participants in the entire exercise included essays representing all four quality points on the RTPET holistic rating scale.

Opening Paragraph Sophistication

The first independent variable, sophistication of rhetorical strategy in opening paragraph, was intended to represent one obvious indicator of writer rhetorical competence. It had two levels. The less sophisticated introductory rhetorical strategy used in this study was characterized by an immediate, first-sentence presentation of the thesis, one which functioned as a direct answer to the assigned topic question. The sample introductory paragraph that follows is a response to the RTPET topic, "Watching the 'soaps' has become an American pastime. Why are these television shows so popular?"

> "Watching daytime soap operas has become a favorite pastime for many Americans. The reason so many people like soap operas is that they offer us something beyond the ordinary world we live in. These shows are filled with attractive stars, excitement, danger, adventure, and of course love. Almost everyone, young and old, watches the soaps, even some very well educated people. Three generations of my family (my grandmother, my mother, and I) watch General Hospital every day. Soap opera fans watch these shows for three reasons: they enjoy the lowbrow entertainment, they relate the events to their own lives and, in a strange way, they even find the shows educational."

The more sophisticated level of introductory strategy used in this study is characterized by what Sheridan Baker (1998) calls a "funnel" introduction: one which takes advantage of the "inevitable psychology of interest" by placing the specific thesis of the essay at or near the end of the first paragraph, after beginning more broadly on the topic but moving purposefully and

logically toward the thesis (pp. 18-19). The following introductory paragraph, a response to the same RTPET topic presented above, has the same number of words and sentences as the previous introduction. However, it is relatively more sophisticated in terms of its broader, less direct opening followed by a purposeful narrowing to the thesis:

> "At home, in college activity centers, even in the waiting rooms of auto repair shops, many Americans' eyes are glued to television screens every weekday afternoon. They "shush" anyone who has the bad judgment to speak while the show is on. What in the world are they so absorbed in? It's not the latest news of Mideast violence or the ups and downs of Wall Street. Rather, they are satisfying an addiction to "their stories"– soap operas. Soap opera fans watch these shows for three reasons: they enjoy the lowbrow entertainment, they relate the events to their own lives and, in a strange way, they even find the shows educational."

Numerous widely used college writing texts (Baker, 1998; Eggers, 1998; Fawcett, 2004; Langan, 2001; Lunsford and Connors, 1995) describe the immediate presentation of the thesis as less rhetorically effective or less sophisticated than introductions which preface the thesis with material designed to engage the reader's interest.

Organizational Consistency

The second independent variable, discrepancies between organizational cues in the opening paragraph and organizational fulfillment instantiated in the subsequent paragraphs, had three levels. The first level reflected consonance between points one, two, and three of the introductory organizational frame and the main ideas in the three body paragraphs. In other words, organizational level 1 was represented when the opening paragraph promised to discuss points A, B, and C, then paragraph two was about A, paragraph three was about B, and

paragraph four was about C. The second level of organizational consistency was characterized by dissonance between point one of the introductory organizational framework and the main idea of paragraph two. That is, if the introductory paragraph promised that the first point to be discussed would be topic A, but paragraph 2 was about topic B, then organizational level 2 was thereby represented. The third level of organization reflected dissonance between point three of the introductory frame and the main idea of paragraph four.

In each case the main idea of body paragraphs was made obvious and its position was held constant across essays. In every body paragraph of every experimental essay, the main idea was instantiated in an explicit topic sentence located at the beginning of the paragraph.

Location of Error Density

The third independent variable had two levels, early versus late density of error. Density of error was operationalized here as in Study 1. That is, each experimental version of each original essay was modified to have early or late density of error, using the selected groupings of the error types found in samples of RTPET essays, as described in Chapter 3. (See Table 3.2 for the types and frequencies of errors used in the error infusion.) Half of the experimental essays concentrated a selected group of errors characteristic of failing essays in the first half of the essay, while a much smaller number of selected errors representative of minimally passing essays were dispersed throughout the second half of the essay. An equal number of experimental essays inverted this pattern: fewer errors, representative of essays of minimal passing quality, in the first half of the essay; more errors, representative of failing essays, in the second half. Word counts were used to determine the midpoint of each essay. Errors were dispersed throughout each half of every essay according to the high and low density scheme. All 24 experimental essays controlled density of error in similar fashion.

Paragraph Sequence

Raters rendered composition quality judgments at the end of each paragraph. They were asked to rate the quality of the entire essay based on what they had read to each point. Following the final paragraph, they were asked to offer their assessment of the entire essay. In each case, essays were composed of five paragraphs. Thus paragraph sequence, from first paragraph to fifth paragraph, captured the unfolding sequence of raters' serial judgments.

Replication

The fifth independent variable, replication, had two levels (replication one and replication two). To determine whether the findings of the study could be extrapolated to the general case of extemporaneous student essays written under standardized conditions, two sets of 12 experimental essays, each set representing the 12 different treatment conditions, were included in the study. The 24 essays selected to be used as experimental essays were nonsystematically distributed to the two replication conditions prior to being modified to fit the 12 treatment conditions. This variable was of no extrinsic interest other than to help establish generalizability of findings across specific essays.

Procedures

The study used a computerized, online presentation of essays one paragraph at a time, with raters required to report a tentative judgment of writing quality before seeing each new paragraph within an essay. This method ensured sequential collection of raters' impressions of writing quality in real time. It allowed examination of the cumulative formation of holistic raters' judgments of writing quality within essays representing each of the 12 treatment conditions. It made possible comparisons of quality ratings within essays at each of five different points.

Each of the raters was asked to read the entire set of essays, including the 24 experimental and 18 filler essays, in a laboratory context using online computer presentation and sequential paragraph-by-paragraph reporting of scores for each essay. Prior to beginning the project, two training sessions were conducted. First, raters were led through a traditional practice scoring session with paper essays and a simple holistic score for each essay. The purpose of this session was to familiarize raters with benchmark essays reflecting each of the four points on the RTPET holistic scoring scale. Next, raters were trained to use the computerized essay presentation and score reporting system. After all participants reported comfort with the online system, the raters were allowed to begin the formal scoring session.

Each rater was exposed to two replications of all 12 treatment conditions in the 24 experimental essays, which were dispersed among the filler essays in one of four randomized orders of presentation.

The texts were presented in cumulative fashion to raters, beginning with just the first paragraph and then adding each subsequent paragraph in sequence, on individual computer screens. As they proceeded through each essay, raters were prompted to indicate, by means of a mouse and an online scoring grid, their tentative judgment of the quality of the cumulative portion of the essay seen up to the point of each scoring prompt. Raters were prompted to report their sequential judgments via a software program which presented them the next paragraph only after they had indicated a judgment of the full portion of the essay seen up to that point. A scoring grid appeared at the bottom of the computer screen with each new paragraph displayed onscreen. Above each grid, raters were asked to report their rating of the portion of the essay they had read up to that point. (To see the scoring grid as it appeared to raters online, see Figure 4.1. Appendix B illustrates how an entire essay was presented to raters paragraph by paragraph).

To allow for recursive reading, the entire preceding text of the essay was available to raters via a scroll bar as each new paragraph was presented. After raters had been presented with the fifth paragraph of each essay, they were prompted to report their final rating of the essay.

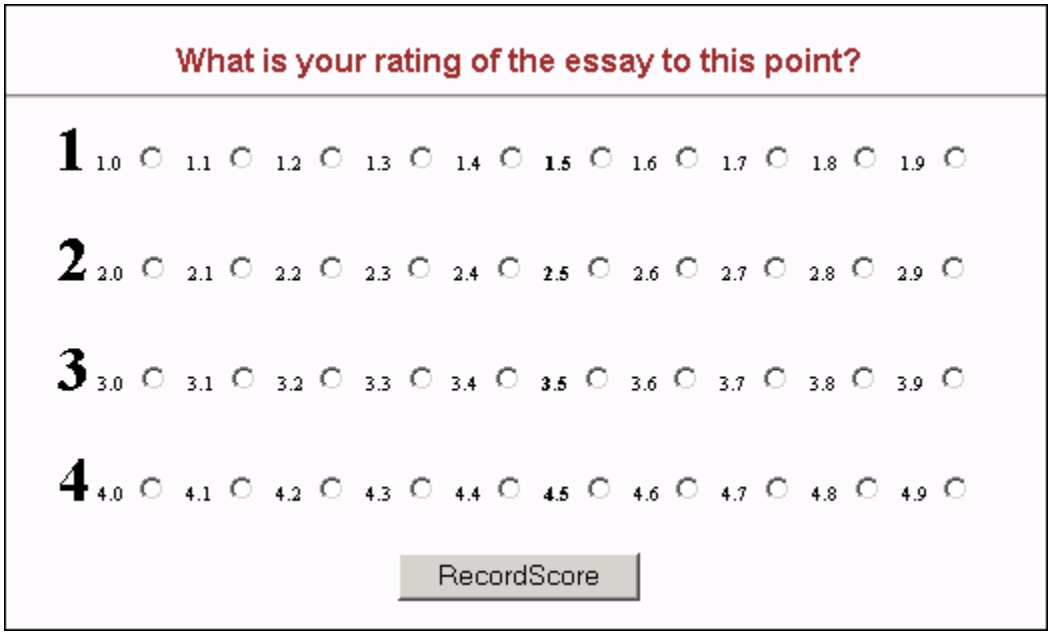


Figure 4.1: Online Scoring Grid (40-point scale)

The rating scale was based on the four-point holistic scale used in the RTPET. However, the scale was expanded to become a forty-point scale, with the increased number of data points allowing raters to give more finely calibrated ratings of the quality of each chunk of text, allowing a more subtle examination of variations in the unfolding development of raters' judgments about writing quality**.** Each of the four quality points in the RTPET scale was expressed along the following ranges: from 1.0 to 1.9, 2.0 to 2.9, 3.0 to 3.9, and 4.0 to 4.9.

Analysis

To prevent any unanticipated confounding with order of reading essays, essays were presented to raters in one of four random and uninterpretable orders. The first analysis, therefore, was to ascertain the impact of order of presentation, random though it was, on average

essay ratings. Next, a five-way repeated measures ANOVA was run. The five factors were opening paragraph sophistication at two levels, organizational consistency at three levels, location of error density at two levels, paragraph sequence at five levels, and replication at two levels. All of these factors were repeated measures. That is, each participant encountered both levels of opening paragraph sophistication, all three levels of organizational consistency and both levels of error density location. They encountered all 12 combinations of these factors in each of the two replications (that is, in two different essays). Moreover, each participant rendered a composition quality judgment at the conclusion of each of the five paragraphs that comprised each of the 24 essays he or she read. The sole rationale for the replication factor was to ensure that results could generalize beyond a single essay. Other than its function to ensure generalizability, the replication factor itself was of no theoretic interest. The sole ANOVA effect of genuine interest in this study was the four-way interaction between the manipulated stylistic factors--that is, between the error, organizational, and introductory factors–and sequence of rating (i.e., paragraph sequence). It is within this interaction that answers to the research questions resided. That data analysis proceeded by conducting a priori nonorthogonal comparisons of interest (Dunn's multiple comparisons, that is, Bonferroni t-tests).

Results

The 120 cell means resulting from the crossing of the five independent variables appear in Appendix C. A summary table for the full ANOVA appears in Appendix D. As shown in Appendix D, the four-way interaction of interest, the interaction between introductory strategy (sophisticated "funnel" vs. less sophisticated direct statement), location of error density (low density followed by high vs. high error density followed by low), organizational consistency (fully consistent vs. violation at paragraph 2 vs. violation at paragraph 4), and temporal sequence

of rating (after reading paragraph 1, paragraph 2, paragraph 3, paragraph 4, or at the conclusion of the essay) attained statistical significance and accounted for about a quarter of all the variance in the ratings ($F_{8,88} = 3.48$, $p < .005$., $eta^2 = .24$). The five way interaction that also included replication (that is, two essays were used to represent each combination of the four factors of interest) did not attain statistical significance ($F_{8,88} = 1.14$, $p = .346$). This means that the four-way interaction could be understood to have generalized across different essay texts. Replication, therefore, could be averaged across (ignored) in the subsequent analyses. Because any lower-order main or interaction effect was modified by the statistically significant 4-way interaction, and because the higher order 5-way interaction was not statistically significant, it was justified to limit the statistical analysis to examining cell mean comparisons comprising the four-way interaction among the factors of interest in this study.

To investigate pairwise cell comparisons within this interaction, preplanned nonorthogonal contrasts (Dunn's multiple comparisons [Bonferroni t-tests] with a family-wise error rate of .05) tested all 240 pairwise contrasts within simple effects. Sixty-nine pairwise contrasts exceeded the critical value (> 1.996) for statistical significance. These contrasts are reported in Table 4.1.

Table 4.1

Significant Pairwise Contrasts Between Cell Means Within 4-Way Interaction: Introductory Strategy x Organizational Consistency x Location of Error x Paragraph

Pairwise Contrast	Cell Means	Difference in Cell Means
P2 I1 O1 E1 → P2 I1 O1 E2	20.0417 → 24.0417	-4.0000
P3 I1 O1 E1 → P3 I1 O1 E2	20.9167 → 23.5000	-2.5833
P1 I1 O2 E1 → P1 I1 O2 E2	18.3750 → 23.3750	-5.0000
P2 I1 O2 E1 → P2 I1 O2 E2	17.5833 → 21.0833	-3.5000
P3 I1 O2 E1 → P3 I1 O2 E2	18.3750 → 23.4167	-5.0417
P4 I1 O2 E1 → P4 I1 O2 E2	18.5417 → 22.9167	-4.3750
P5 I1 O2 E1 → P5 I1 O2 E2	18.7083 → 21.9583	-3.2500
P1 I1 O3 E1 → P1 I1 O3 E2	19.4167 → 21.5417	-2.1250
P2 I2 O1 E1 → P2 I2 O1 E2	19.4167 → 23.2500	-3.8333

P3 I2 O1 E1 → P3 I2 O1 E2	20.0417 → 23.8750	-3.8333
P1 I2 O2 E1 → P1 I2 O2 E2	20.0417 → 23.9167	-3.8750
P2 I2 O2 E1 → P2 I2 O2 E2	19.7917 → 23.2083	-3.4166
P3 I2 O2 E1 → P3 I2 O2 E2	20.7500 → 23.2500	-2.5000
P4 I2 O2 E1 → P4 I2 O2 E2	20.7083 → 23.4583	-2.7500
P1 I2 O3 E1 → P1 I2 O3 E2	23.0833 → 25.1667	-2.0834
P2 I2 O3 E1 → P2 I2 O3 E2	21.4583 → 26.0833	-4.6250
P3 I2 O3 E1 → P3 I2 O3 E2	21.7083 → 26.5833	-4.8750
P4 I2 O3 E1 → P4 I2 O3 E2	22.8333 → 25.4583	-2.6250
P5 I2 O3 E1 → P5 I2 O3 E2	22.2500 → 25.5417	-3.2917
P1 O1 E1 I1 → P1 O1 E1 I2	24.0833 → 21.6250	-2.4583
P2 O2 E1 I1 → P2 O2 E1 I2	17.5833 → 19.7917	-2.2084
P3 O2 E1 I1 → P3 O2 E1 I2	18.3750 → 20.7500	-2.3750
P4 O2 E1 I1 → P4 O2 E1 I2	18.5417 → 20.7083	-2.1666
P5 O2 E1 I1 → P5 O2 E1 I2	18.7083 → 20.8333	-2.1250
P1 O3 E1 I1 → P1 O3 E1 I2	19.4167 → 23.0833	-3.6666
P4 O1 E2 I1 → P4 O1 E2 I2	23.7083 → 21.0000	2.7083
P2 O2 E2 I1 → P2 O2 E2 I2	21.0833 → 23.2083	-2.1250
P1 O3 E2 I1 → P1 O3 E2 I2	21.5417 → 25.1667	-3.6250
P2 O3 E2 I1 → P2 O3 E2 I2	21.9583 → 26.0833	-4.1250
P3 O3 E2 I1 → P3 O3 E2 I2	21.6667 → 26.5833	-4.9166
P4 O3 E2 I1 → P4 O3 E2 I2	21.5833 → 25.4583	-3.8750
P5 O3 E2 I1 → P5 O3 E2 I2	21.1667 → 25.5417	-4.3750
P1 I1 O1 E1 → P2 I1 O1 E1	24.0833 → 20.0417	4.0416
P1 I1 O1 E1 → P3 I1 O1 E1	24.0833 → 20.9167	3.1666
P1 I1 O3 E1 → P4 I1 O3 E1	19.4167 → 21.6667	-2.2500
P1 I1 O3 E1 → P5 I1 O3 E1	19.4167 → 21.5417	-2.1250
P1 I1 O2 E2 → P2 I1 O2 E2	23.3750 → 21.0833	2.2917
P2 I1 O2 E2 → P3 I1 O2 E2	21.0833 → 23.4167	-2.3334
P1 I2 O1 E1 → P2 I2 O1 E1	21.6250 → 19.4167	2.2083
P2 I2 O1 E1 → P4 I2 O1 E1	19.4167 → 22.5000	-3.0833
P2 I2 O1 E1 → P5 I2 O1 E1	19.4167 → 22.6667	-3.2500
P3 I2 O1 E1 → P4 I2 O1 E1	20.0417 → 22.5000	-2.4583
P3 I2 O1 E1 → P5 I2 O1 E1	20.0417 → 22.6667	-2.6250
P1 I2 O1 E2 → P4 I2 O1 E2	23.2083 → 21.0000	2.2083
P2 I2 O1 E2 → P4 I2 O1 E2	23.2500 → 21.0000	2.2500
P3 I2 O1 E2 → P4 I2 O1 E2	23.8750 → 21.0000	2.8750
P1 I1 E1 O1 → P1 I1 E1 O2	24.0833 → 18.3750	5.7083
P2 I1 E1 O1 → P2 I1 E1 O2	20.0417 → 17.5833	2.4584
P3 I1 E1 O1 → P3 I1 E1 O2	20.9167 → 18.3750	2.5417
P4 I1 E1 O1 → P4 I1 E1 O2	22.2917 → 18.5417	3.7500
P5 I1 E1 O1 → P5 I1 E1 O2	22.2083 → 18.7083	3.5000
P1 I1 E1 O1 → P1 I1 E1 O3	24.0833 → 19.4167	4.6666
P2 I1 E1 O2 → P2 I1 E1 O3	17.5833 → 20.7500	-3.1667
P3 I1 E1 O2 → P3 I1 E1 O3	18.3750 → 20.5000	-2.1250
P4 I1 E1 O2 → P4 I1 E1 O3	18.5417 → 21.6667	-3.1250

P5 I1 E1 O2 → P5 I1 E1 O3	18.7083 → 21.5417	-2.8334
P2 I1 E2 O1 → P2 I1 E2 O2	24.0417 → 21.0833	2.9584
P4 I1 E2 O1 → P4 I1 E2 O3	23.7083 → 21.5833	2.1250
P5 I1 E2 O1 → P5 I1 E2 O3	23.4167 → 21.1667	2.2500
P2 I2 E1 O1 → P2 I2 E1 O3	19.4167 → 21.4583	-2.0416
P1 I2 E1 O2 → P1 I2 E1 O3	20.0417 → 23.0833	-3.0416
P4 I2 E1 O2 → P4 I2 E1 O3	20.7083 → 22.8333	-2.1250
P2 I2 E2 O1 → P2 I2 E2 O3	23.2500 → 26.0833	-2.8333
P3 I2 E2 O1 → P3 I2 E2 O3	23.8750 → 26.5833	-2.7083
P4 I2 E2 O1 → P4 I2 E2 O3	21.0000 → 25.4583	-4.4583
P5 I2 E2 O1 → P5 I2 E2 O3	22.0000 → 25.5417	-3.5417
P3 I2 E2 O2 → P3 I2 E2 O3	23.2500 → 26.5833	-3.3333
P4 I2 E2 O2 → P4 I2 E2 O3	23.4583 → 25.4583	-2.0000
P5 I2 E2 O2 → P5 I2 E2 O3	22.4167 → 25.5417	-3.1250

The research questions, however, pertain only to the 120 contrasts that compare ratings within combinations of the stylistic factors across paragraphs. That is, in keeping with the stated objective to examine how composition ratings emerge over real time as additional text becomes available to readers, this analysis focuses only on those simple effects that compare ratings across paragraphs. These are repeated measure comparisons in which raters are essentially compared against their own unfolding scoring**.** In addition, the very practical question of how changes in organizational consistency and error density affect final paper ratings warrants examination of one more simple effect: the comparison of final (paragraph 5) ratings across each of the 12 combinations of introductory sophistication, error density, and organizational consistency.

In each set of five cell means embedded within the analyses that follow, every cell mean is labeled with this sequence of factor-positions: I = introductory condition, E = location of error density condition, O = organization condition, and P = paragraph condition. Therefore, the symbol M_{IEOP} represents the mean rating for the I^{th} level of introductory strategy (I = 1,2), the E^{th} level of location of error density (E = 1,2), the O^{th} level of organizational consistency (O = 1,2,3), and the P^{th} level of paragraph sequence (P = 1,2,3,4,5).

We first consider cross-paragraph contrasts within the combination of factors that we would hypothesize to yield the worst initial impression on raters: less sophisticated introductory strategy; high error density in the first half, and early violation of organizational scheme (that is, organizational violation appearing in paragraph 2). As the set of five cell means below indicates, there are no significant changes in ratings. The ratings start off low after the first paragraph, and they never rise, even, for example, when the error rate drops precipitously at the midway point of the essay. Apparently the initial negative impression of writing quality was sufficiently powerful that it held firm in spite of improved command of mechanical correctness later in the essay.

$M_{I1,E1,O2,P1}$	$M_{I1,E1,O2,P2}$	$M_{I1,E1,O2,P3}$	$M_{I1,E1,O2,P4}$	$M_{I1,E1,O2,P5}$
18.38	17.58	18.38	18.54	18.71

In the similar configuration (less sophisticated introduction and high error density initially) but with the organizational violation withheld until paragraph 4, ratings were higher at paragraph 4 and paragraph 5 relative to the composition rating given after reading just paragraph 1. In this case, it appears that the reduction in error rate later in the essay overwhelmed any negative evaluation that might have been due to the organizational violation late in the essay.

$M_{I1,E1,O3,P1}$	$M_{I1,E1,O3,P2}$	$M_{I1,E1,O3P3}$	$M_{I1,E1,O3P4}$	$M_{I1,E1,O3P5}$
19.42	20.75	20.50	21.67	21.54

In the parallel set of conditions (high error density initially and less sophisticated introduction) but with no organizational violation whatsoever in the essay, the quality rating at the first paragraph was higher than at either paragraph 2 or paragraph 3. This may be an anomaly of a particularly high rating of the first paragraph in that condition. (Appendix C indicates that it was rated significantly higher than the first paragraphs in either of the two

preceding configurations, even though all three first paragraphs were identical in terms of treatment condition.)

$M_{I1,E1,O1,P1}$	$M_{I1,E1,O1,P2}$	$M_{I1,E1,O1,P3}$	$M_{I1,E1,O1,P4}$	$M_{I1,E1,O1,P5}$
24.08	20.04	20.92	22.29	22.21

Consider now the cross-paragraph contrasts within the combination of factors which we would expect to yield the best first impression: low density of error followed by higher density, sophisticated "funnel" introduction, and no violation of organizational premise. Here a significant drop in ratings is evident when the higher density of error is introduced at the halfway mark of the essay. Paragraph 4 elicited lower ratings than had been given at any of the preceding three, which did not differ among themselves. In this condition, when raters encountered the increased error rate at paragraph 4, their negative reactions were not mitigated by the stronger opening paragraphs of this essay.

$M_{I2,E2,O1,P1}$	$M_{I2,E2,O1,P2}$	$M_{I2,E2,O1,P3}$	$M_{I2,E2,O1,P4}$	$M_{I2,E2,O1,P5}$
23.21	23.25	23.88	21.00	22.00

However**,** in the similar configurations in which lower error density was encountered first and the introductory paragraph strategy was sophisticated, but in which organizational violations occurred either at paragraph 2 or at paragraph 4, ratings did not change at all throughout the readings. That is, in these cases raters were not significantly moved from the impressions they formed in paragraph 1, even when they encountered those organizational violations and even when they encountered the increase in error rate at the midpoint of the essay. It is almost as if the organizational violations protected the ratings from plummeting due to the mid-essay onslaught of effort.

Organizational violation at paragraph 2:

$M_{I2,E2,O2,P1}$	$M_{I2,E2,O2,P2}$	$M_{I2,E2,O2,P3}$	$M_{I2,E2,O2,P4}$	$M_{I2,E2,O2,P5}$
23.92	23.21	23.25	23.46	22.42

Organizational violation at paragraph 4:

$M_{I2,E2,O3,P1}$	$M_{I2,E2,O3,P2}$	$M_{I2,E2,O3,P3}$	$M_{I2,E2,O3,P4}$	$M_{I2,E2,O3,P5}$
25.17	26.08	26.58	25.46	25.54

Several permutations lie between those extremes of hypothesized positive expectations and negative expectations. First, within the sophisticated introduction strategy, several conditions presented a high density of error, switching to lower density only in the second half of the essay. Within that combination, when the promised organizational structure was violated at paragraph 2, as well as when it was violated at paragraph 4, there was no significant difference in judged composition quality as the paper progressed. Even when the density of errors decreased at the midpoint of the essay, subsequent ratings did not reflect that improvement.

Organizational violation at paragraph 2:

$M_{I2,E1,O2,P1}$	$M_{I2,E1,O2,P2}$	$M_{I2,E1,O2,P3}$	$M_{I2,E1,O2,P4}$	$M_{I2,E1,O2,P5}$
20.04	19.79	20.75	20.71	20.83

Organizational violation at paragraph 4:

$M_{I2,E1,O3,P1}$	$M_{I2,E1,O3,P2}$	$M_{I2,E1,O3,P3}$	$M_{I2,E1,O3,P4}$	$M_{I2,E1,O3,P5}$
23.08	21.46	21.71	22.83	22.25

When there were no violations of the organization that was promised in paragraph 1, however, improvement at the latter part of the essay was discerned. In this treatment combination, the drop in error density at the midpoint of the essay resulted in higher ratings of the composition than was the case after reading paragraph 1. And the final essay rating, following the reading of paragraph 5, was significantly higher than the ratings given after reading paragraph 2 and paragraph 3. For no apparent reason there was also a statistically significant drop in the composition rating between the first two paragraphs.

$M_{I2,E1,O1,P1}$	$M_{I2,E1,O1,P2}$	$M_{I2,E1,O1,P3}$	$M_{I2,E1,O1,P4}$	$M_{I2,E1,O1,P5}$
21.63	19.42	20.04	22.50	22.67

Within the less sophisticated introductory strategy (that is, a bald statement of the essay thesis), when readers encountered first a low density of errors and only later a higher density, initial impressions of the composition never changed across the final paragraph when (a) there were no violations of the promised organization, nor when (b) organizational inconsistency was introduced midway through the essay.

No organizational violation:

$M_{I1,E2,O1,P1}$	$M_{I1,E2,O1,P2}$	$M_{I1,E2,O1,P3}$	$M_{I1,E2,O1,P4}$	$M_{I1,E2,O1,P5}$
25.08	24.04	23.50	23.71	23.42

Organizational violation at paragraph 4:

$M_{I1,E2,O3,P1}$	$M_{I1,E2,O3,P2}$	$M_{I1,E2,O3,P3}$	$M_{I1,E2,O3,P4}$	$M_{I1,E2,O3,P5}$
21.54	21.96	21.67	21.58	21.17

However, within this general configuration (that is, unsophisticated introduction and high error rate only in the second half of the essay), when organizational inconsistency was introduced at paragraph 2, reader ratings did shift significantly. They declined when the inconsistency was encountered in paragraph 2, and then they rose back up when paragraph 3 did deliver on the topical expectation promised in the introduction. In none of these three treatment combinations, however, did raters shift their judgments when they encountered the higher density of error after the midpoint of the essay.

$M_{I1,E2,O2,P1}$	$M_{I1,E2,O2,P2}$	$M_{I1,E2,O2,P3}$	$M_{I1,E2,O2,P4}$	$M_{I1,E2,O2,P5}$
23.38	21.08	23.42	22.92	21.96

Analysis of Whole Essays (Final Paragraph Ratings)

In addition to examining the unfolding evaluations of essays by comparing paragraph ratings within each of the treatment combinations, as above, the final paragraph or holistic essay

ratings were compared within simple effects. Thus, for example, when errors were introduced early and reduced later and the introduction was sophisticated, final paragraphs were compared among the three levels of organizational inconsistency. Eighteen such contrasts were calculated. Nine proved statistically significant, and these are presented in Table 4.2. As Table 4.2 indicates, five of these simple effects contrasts involved comparisons across levels of organizational inconsistency. Essays that began with simple introductions, had error in early rather than in later sections, and maintained organizational consistency were rated higher than the otherwise similar essays that had subsequent organizational violations early in the essay (paragraph 2). Essays that had simple introductions, had error appearing in later rather than in earlier sections, and were organizationally consistent were rated higher than otherwise similar essays that manifested organizational inconsistency later in the essay (paragraph 4). Essays that began with simple introductions, had error in early rather than in later sections, and manifested organizational inconsistency late in the essay (paragraph 4) were rated higher than otherwise similar essays that manifested organizational inconsistency relatively earlier (paragraph 2). Essays that began with sophisticated introductions, had error appearing in later sections rather than in earlier sections, and manifested organizational inconsistency late in the essay (paragraph 4) were rated higher than the otherwise similar essays with inconsistency manifested earlier (paragraph 2).

Table 4.2

Significant Pairwise Contrasts Between Cell Means of Overall Essay Ratings Within Significant 4-Way Interaction (Introductory Strategy x Organizational Consistency x Location of Error x Paragraph)

Pairwise Contrast	Cell Means	Difference in Cell Means
P5 I1 O2 E1 → P5 I1 O2 E2	18.7083 → 21.9583	-3.2500
P5 I2 O3 E1 → P5 I2 O3 E2	22.2500 → 25.5417	-3.2917
P5 O2 E1 I1 → P5 O2 E1 I2	18.7083 → 20.8333	-2.1250
P5 O3 E2 I1 → P5 O3 E2 I2	21.1667 → 25.5417	-4.3750
P5 I1 E1 O1 → P5 I1 E1 O2	22.2083 → 18.7083	3.5000
P5 I1 E1 O2 → P5 I1 E1 O3	18.7083 → 21.5417	-2.8834
P5 I1 E2 O1 → P5 I1 E2 O3	23.4167 → 21.1667	2.2500
P5 I2 E2 O1 → P5 I2 E2 O3	22.0000 → 25.5417	-3.5417
P5 I2 E2 O2 → P5 I2 E2 O3	22.4167 → 25.5417	-3.1250

The pattern thus far explicated indicates that the more of the essay read without encountering organizational inconsistency, the higher it was rated. However, this pattern was reversed in one of the final paragraph comparisons. When essays began with a more sophisticated introduction, had error appearing in later segments rather than in earlier segments, and manifested no organizational inconsistency, they were rated more poorly than otherwise similar essays which manifested organizational inconsistency later in the essay (paragraph 4). This latter contrast defies principled explanation.

One further pair of final essay comparisons revealed the advantage of sophisticated introductory strategies on ratings. When organizational inconsistency occurred early in the essay and error also occurred earlier rather than later in the essay, and a sophisticated funnel introduction was used, the essay was rated higher than the otherwise similar essays with the less sophisticated introduction. Similarly, when organizational inconsistency occurred later in the essay, error density was greater in the later than in the earlier segments, and a sophisticated introduction was used, the essay was rated higher than otherwise similar essays with simple introductions.

Two additional contrasts on final ratings revealed the advantage of late-occurring error density. When an essay had a simple introduction, organizational inconsistency occurred early rather than later, and error predominated in later sections rather than in earlier sections, the essay received a higher rating than otherwise similar essays in which error occurred earlier rather than later. And when an essay had a sophisticated introduction, organizational inconsistency occurred later rather than earlier in the essay, and error occurred later rather than earlier, the essay was rated higher than otherwise similar essays in which error appeared earlier.

In sum, with the exception of one anomalous comparison between cell means, these contrasts between final paragraph ratings indicated an advantage for organizational consistency, or at least late-onset of organizational inconstancy, as compared with early-onset organizational inconsistency. The findings also indicated an advantage for later occurring mechanical errors over essays that had dense errors toward the beginning. Finally, this set of findings indicated an advantage for essays which are mechanically correct to start off with (even though they may regress in their closings), as compared to essays with dense mechanical errors at the beginning but few at the end.

CHAPTER 5

STUDY 3

This chapter describes the participants, the test instruments used, the methods for collecting the data, the procedures for analyzing the data, and the analysis and results for Study 3. Study 3 is motivated by claims—largely unexamined—that portfolio assessment constitutes a more fair method of assessing student writing proficiency than single high-stakes writing samples (Larson, 1996). To better evaluate these claims about the fairness of portfolio assessment, it is important to know the degree to which raters are capable of judging portfolios holistically, unhampered by extraneous factors like the position of each essay within the portfolio, factors that might cause one component of the portfolio to exert disproportionate impact on raters' perceptions. This study extended to a portfolio assessment context the investigation of how holistic raters' judgments develop. It investigated the question whether evaluative impressions formed by exposure to essays of lower or higher levels of quality, relative to the other essays within the portfolio, influence raters' final judgments about portfolios. This study also examined the effect of varying the position of the dissonant essay within the portfolio on the overall ratings assigned by raters to portfolios.

The following are the research questions which were investigated in Study 3:

RQ3.0 Given a portfolio consisting of three writing samples, does the inclusion of a writing sample of higher or lower quality than the other two samples affect the ratings of the portfolio as a whole?

RQ3.1 Does the position of the dissonant writing sample in relation to the other two samples influence the rating of the portfolio as a whole? That is, does either the primacy or recency of a dissonant quality writing sample within a portfolio affect the overall evaluation of that portfolio?

Method

Study 3 investigated the rating behaviors of readers evaluating portfolios holistically in a laboratory setting. All raters were asked to evaluate the same set of portfolios, using a scoring rubric adapted from the RTPET to fit portfolio assessment. Raters were asked to assign a holistic rating to each portfolio. In this study, two independent variables were of interest – dissonance level (high or low) of the discrepant-quality essay and position of extemporaneous essay within the portfolio (first, middle, or last). The portfolio ratings were the sole outcome variable.

Participants

The participants in this study were the same participants as in Study 2. All participants were experienced instructors of freshman composition at a college or university within the University System of Georgia and experienced RTPET raters. All had participated in multiple state-conducted training sessions to prepare them to evaluate individual student essays written for the Georgia RTPET using the Regents' Testing Program holistic scoring guidelines. Participants were financially compensated for their involvement in the study at a rate identical to the daily rate of remuneration received by RTPET raters.

Portfolios/Essays

Portfolios were constructed so as to represent, respectively, two and three levels of two independent variables: quality of extemporaneous essay (higher than or lower than the other

essays in the portfolio) and placement of target essay in the three-essay portfolio (first, middle, and end positions). Although the particular essays selected for use in this study were not of direct interest, essay also functioned as an independent variable at six different levels, since the experimental essays were necessarily different so as to prevent raters from encountering the same essay more than once.

Fifteen writing portfolios were developed for use in this research project. To eliminate handwriting as a confounding variable, all portfolios were word processed. Each portfolio consisted of three essays written in response to topics used in the Georgia RTPET. Most of these essays were selected from a pool of authentic RTPET essays written in a previous administration of the RTPET. A few of the essays needed, however, could not be retrieved from existing test archives. They were therefore either written by university students under standardized conditions similar to those used in the RTPET and then modified by the researchers, or were written by the researchers themselves to simulate student essays.

Each portfolio was comprised of three essays represented as having been written by one freshman English student. Four essay topics selected from the list of topics used for the RTPET were used. In each portfolio, one essay was represented as having been written in practice RTPET conditions, which required students to write extemporaneously, independently, and under proctor supervision for no more than an hour. It was anticipated that raters might give particular credence to an essay written under supervision and with no assistance, as a more accurate indicator of student writing proficiency than an untimed writing sample produced under unsupervised conditions and/or incorporating teacher feedback. The other two essays, also expository, were therefore represented as having been written as course assignments outside of class, over time, with cycles of feedback and revision. However, these essays were comparable

in length to the extemporaneous essays with which they were grouped; no essay varied more than twenty-five per cent in word count from the other essays with which it was grouped.

Prior to inclusion in the study, each essay used in this project was predetermined to reflect one of three quality ratings (1, 2, or 3) on the four-point holistic scale used in the Regents' Testing Program Essay Test. Because of the infrequency of essays rated at the highest quality rating (level 4) in the naturalistic context of the RTPET, no essays rated at the "4" quality score were used in this research project. Those essays that were actual RTPET essays had been rated in the naturalistic scoring procedure used in a previous administration of the Regents' Testing Program; the ratings assigned to them in those circumstances were simply accepted for use in this study. The additional several essays that were modified or written for the study were submitted to veteran RTPET raters who had maintained records of meeting or exceeding RTPET standards for interrater agreement, and the consensus quality ratings of those essays were determined by these raters (via agreement of the two raters initially consulted, or two out of three in instances where the first two raters reached a split decision and a third rater was consulted). These ratings were used as measures of the quality of all the essays included in the study.

Nine portfolios in the study contained three essays of homogeneous quality, three each at levels "1," "2," and "3." These homogeneous portfolios served as benchmarks against which the portfolios containing dissonant quality essays could be arrayed.

The six experimental portfolios containing essays of dissonant quality were constructed in such a way as to allow inferences about the relative impact of the extemporaneous essay versus the untimed, revised essays. In half the portfolios, then, the designated extemporaneous essay was of higher quality than the two untimed essays. These are referred to as "high dissonant" portfolios. In other portfolios, the designated extemporaneous essay was of lower

quality than the two untimed essays. These are referred to as "low dissonant" portfolios. Dissonance among essays was thus treated as an independent variable with two levels: dissonant higher quality and dissonant lower quality extemporaneous essays. Each experimental portfolio was constructed so as to make the quality of the extemporaneous essay dissonant with the quality of the other two essays. In three portfolios containing dissonant higher quality essays, the extemporaneous essay was of higher quality (level 3 on the RTPET scale) than the other two essays (level 1); in three portfolios containing a dissonant lower quality essay, the extemporaneous essay was of lower quality (level 1 on the RTPET scale) than the others (level 3).

To determine whether the placement of the dissonant extemporaneous essay would affect the rating of the overall portfolio, the order in which the designated extemporaneous essay was presented to raters was also treated as an independent variable. In one third of the experimental portfolios the designated extemporaneous essay was the first one which raters encountered. In another third, raters encountered the designated extemporaneous essay in middle position, and in the remaining third the extemporaneous essay appeared as the third and last item in the portfolio. Essay placement was crossed with the two levels of dissonant quality (high and low), resulting in the six experimental portfolios.

Although not of primary interest in this study, each of the nine homogeneous portfolios (three each at levels 1, 2, and 3) also contained one designated extemporaneous essay appearing in either first, middle, or final position.

Procedures

The raters in this study were informed that this research project was being conducted to pilot test a new system of portfolio assessment that might serve as an alternative to the type of

single-sample test used in the Regents' Testing Program. They were told that this new system had been designed to preserve some of the elements of the RTPET, including the use of the writing genres and topics employed in the RTPET and the inclusion in the portfolio of an essay written in standard RTPET conditions. They were also informed, however, that the assessment procedures had been modified to allow students' work written under more process-oriented conditions to be included in the evaluation process. Thus, they were told that some of the essays they would encounter in each portfolio were standard RTP essays, while others had been produced over time, with the benefit of teacher feedback. Further, they were told that this portfolio evaluation system was being tested to see whether an acceptable level of interrater reliability could be achieved.

Each rater was trained to score portfolios using a rubric (see Appendix E, "Instructions for Scoring Portfolios") which was adapted from "Instructions for Scoring Regents' Testing Program Essays" (see Appendix A) to fit a portfolio assessment rather than the assessment of a single essay. Instructions read, in part,

> "Raters should read each portfolio quickly to gain a general impression of its overall quality. Essays within a portfolio should be read in the order in which they are presented from first to last. However, raters should read quickly to evaluate the entire portfolio holistically and assign one rating to the portfolio as a unit. For this assessment, the raters' task is to assess the quality of the student's writing based on the portfolio as a whole."
>
> "Raters should suspend judgment about the quality of a portfolio until they have finished reading the entire portfolio."

The modified rubric used a four-point holistic scale similar to the RTPET scale to assign a single score to an entire portfolio. Prior to rating experimental portfolios in the actual study, participants were trained in the use of these modified scoring guidelines and the score-reporting system to be used in the study. (See Appendix F to examine the training portfolios and Appendix G to examine score reporting sheets used in training.) As part of this training, participants were asked to rate a set of sample portfolios. They then shared their ratings for each portfolio and discussed their evaluative judgments. Sixty-eight percent of training portfolio ratings were exact agreements among raters. One hundred percent were either exact agreement or else adjacent ratings. For pass-fail decisions, the rate of agreement was 97% among raters. (Note that RTPET standards require only 88% agreement among two out of three raters for each essay.)

To yield data for the analyses of interest in this study, each rater was asked to read and rate the same 15 portfolios independently of other raters. To estimate portfolio order effects (that is, order between portfolios—which was of only nuisance interest), the portfolios were distributed to raters in four randomized and uninterpretable sequences. Three raters were nested in each of these four random orders. In this portfolio assessment session, they were instructed to assign one holistic rating to each portfolio. (The rater scoring form for portfolios is included in Appendix H.)

Analysis

The holistic ratings of the experimental portfolios were subjected to a 2 (dissonant high quality essay, dissonant low quality essay) x 3 (dissonant essay in initial, medial, or final position) x 4 (order of portfolios) mixed factorial ANOVA. Raters were nested in order of presentation of the portfolios.

Results

Order by portfolio type by extemporaneous essay position cell means appear in Appendix I. A summary table for the full ANOVA appears in Table 5.1. As it shows, type of portfolio attained statistical significance and accounted for about 64% of all the variance in the ratings ($F_{4,59} = 55.15$, $p < .05$, $eta^2 = .64$). Order of presentation was statistically significant, although the effect was small ($F = 3.37$, $p < .05$), accounting for just 7.8% of variance in ratings. The position of the extemporaneous dissonant essay within portfolio was not found to be a statistically significant factor influencing the holistic ratings of portfolios (($F_{2,14} = .46$, $p > .05$). There was no statistically significant interaction between order and position of extemporaneous position ($F_{6,59} = .79$, $p > .05$), portfolio type and order ($F_{12,59} = .1.28$, $p > .05$), or position of extemporaneous essay and portfolio type ($F_{8,59} = .49$, $p > .05$). Nor was the three-way interaction between portfolio type, position of extemporaneous essay, and order statistically significant ($F_{24,59} = 1.55$, $p > .05$).

Table 5.1

Summary Table, ANOVA
Introductory Strategy x Organizational Consistency x Location of Error Density

	SS	Df	MS	F	Sig.	Partial Eta^2
Order	2.244	3	.748	3.367	.021	.078
Portfolio Type	49.022	4	12.256	55.150	.000	.648
Extemporaneous Position	.233	2	.117	.525	.593	.009
Order x Portfolio Type	3.422	12	.285	1.283	.237	.114
Order x Extemp Position	1.056	6	.176	.792	.578	.038
Portfolio Type x Extemp Position	.878	8	.110	.494	.859	.032
Order x Portf Type x Extemp Position	8.278	24	.345	1.552	.064	.237
Error	26.667	120	.222			

A post hoc Student-Neuman-Keuls procedure was conducted to ascertain differences among the cell means. The pattern of results it yielded is represented below:

Homogeneous 1 (M=1.19)	High Dissonant (M=1.86)	Homogeneous 2 (M=2.11)	Low Dissonant (M=2.19)	Homogeneous 3 (M=2.81)

The high dissonant portfolio, containing as it did two level 1 essays and an extemporaneous essay of level 3 was judged significantly higher than the homogeneous failing essay, but lower than all the others. The low dissonant portfolio, containing as it did two level 3 essays and an extemporaneous essay of level 1, was judged no different that the homogeneous level 2 portfolio, but lower than the homogeneous level 3 portfolio. The low dissonant portfolio was judged significantly stronger than the high dissonant portfolio.

The position of the means for the two groups of dissonant portfolios between the mean of low homogeneous portfolios and the mean of high homogeneous portfolios, and on either side of the mean for moderate quality homogeneous portfolios, suggests that raters neither focused on the quality of extemporaneous essays, nor did they disregard them. In fact, raters appeared to have assessed the quality of the dissonant portfolios at close to the unweighted average of the quality level of their three component writing samples. Raters did not treat the quality level of the extemporaneous essays as reflective of the overall quality of the portfolio; nor did they discount the extemporaneous essay in making judgments about the overall quality of the portfolio. In other words, the inclusion of a discrepant-quality essay in a portfolio did influence the rating of the portfolio as a whole. The effect was to raise the scores of portfolios with two failing essays and one higher-quality essay and lower the scores of portfolios with two passing essays and one failing essay.

Although order of portfolio presentation did achieve statistical significance as a factor affecting ratings, the strength of the effect ($eta^2 = .078$) was relatively small. Further, with only three raters assigned to each of the four presentation sequences, it seems reasonable to believe

that individual differences among raters may have accounted for this result. At any rate, the orders of portfolio presentation were random, and therefore not interpretable.

To investigate whether the position of the dissonant essay within portfolio affected the pass/fail decision for the portfolio, Chi-square analyses were conducted for each of the two types of experimental portfolios – high dissonant and low dissonant. The analysis for the high dissonant portfolios shows that for these portfolios, the position of the dissonant extemporaneous essay did not affect the overall pass/fail decision for the portfolio ($\chi^2_{2df} = 4.18$; $p > .05$). (The cross-tabulation results of extemporaneous essay position and pass/fail decision for these portfolios are reported in Table 5.2). However, it should be noted that this was not a well conditioned Chi-Square, since the expected value of failures was too small.

Table 5.2

Extemporaneous Essay and Pass/Fail Decision, High Dissonant Portfolios

	Position of Extemporaneous Essay			
	First	Middle	Last	Total
Failing Portfolios	2	1	5	8
Passing Portfolios	10	11	7	28
Total	12	12	12	36

The cross-tabulation of pass/fail decision and position of extemporaneous essay position for low dissonant essay revealed only one failing portfolio (see Table 5.3), so that three cells were well below the rule-of-thumb minimum expected cell mean of 5. It was therefore not possible to draw any conclusion from the Chi-Square analysis of the differential impact of extemporaneous essay position for low dissonant portfolios.

Table 5.3

Extemporaneous Essay and Pass/Fail Decision, Low Dissonant Portfolios

	Position of Extemporaneous Essay			
	First	Middle	Last	Total
Failing Portfolios	0	1	0	1
Passing Portfolios	12	11	12	35
Total	12	12	12	36

Preliminary Discussion of Study 3

The purpose of Study 3 was to determine the degree to which raters' holistic evaluations of portfolios were disproportionately influenced by dissonant quality extemporaneous essays embedded within them, and whether the impression rendered by first-appearing dissonant essays unduly influenced overall judgments of those portfolios. Previous research bearing on this subject gave reason to suppose that raters' holistic perceptions would be largely influenced by the quality of the first elements they encountered in those portfolios (Hamp-Lyons & Condon, 1993). The findings of the present study, however, do not support that conclusion. To the contrary, raters in the present study seemed to adhere to the instructions given them to give equal weight to all components of the portfolios. Where dissonant quality extemporaneous essays appeared, they were not given disproportionate weight. Moreover, the present study failed to reveal any differential impact for the position within the portfolio in which a higher or lower quality essay appeared.

CHAPTER 6

DISCUSSION

Recapitulation of Purpose

The three studies in this dissertation investigated questions which bear upon the issue of construct validity of holistic assessments of both single essays and portfolios as measures of students' writing ability. The objection that most closely motivated the present studies is the contention that single-sample, standardized, rapid holistic assessment procedures miss the intended mark--the accurate measurement of a student's writing ability. If the large-scale holistic scoring of texts is to be defended as a valid method of assessing writing ability, it must stand up to the charges that holistic rating procedures constrain raters to a superficial reading process that causes them to overvalue surface correctness, undervalue other features, and miss textual subtleties that warrant closer, more careful consideration before reaching an assessment decision.

The general focus of the three research projects, which were coordinated as a set of linked studies, was the sequential nature of the evaluative reading process. More specifically, the three studies investigated the real-time dynamics, conscious and subconscious, of the reading process of holistic raters as they assess student texts in high-stakes, large-scale contexts.

This dissertation, then, builds on one area where reading and writing research intersect: the examination of reading processes involved in writing assessment. Freedman and Calfee's (1983) model of reading for assessment of writing quality in particular, supplemented by Frank Smith's (1994) model of expert reading in general, provided a platform for understanding the

reading processes of holistic raters in writing assessments in the field. This project responds to calls for additional research on evaluative reading (e.g., Huot, 1990).

Two of the studies in this dissertation investigated the extent to which sequential impressions, formed as those raters encountered selected features in text, exerted impact on their final judgments about the whole text. Study 1 investigated the rating behaviors of readers evaluating essays in the naturalistic context of a high-stakes writing competency test in the actual statewide university system assessment. Study 2 investigated the unfolding evaluations of raters from paragraph to paragraph as they read these texts from start to finish, in real time, in a computer-based laboratory context. The third study examined questions about how raters' judgments develop as they score writing portfolios holistically. Study 3 investigated the question whether evaluative impressions formed by exposure to essays of lower or higher levels of quality, relative to the other essays within the portfolio, influence raters' final judgments about portfolios. Study 3 also examined the effect of varying the position of the dissonant essay within the portfolio on the overall ratings assigned by raters to portfolios.

The following are the research questions investigated in these studies.

STUDY 1

1.0 To what extent are raters' judgments of an entire essay in a large-scale testing context influenced by the placement of errors in the essay?

STUDY 2

2.0 To what extent are raters' sequential, paragraph-by-paragraph evaluative impressions of an essay in a laboratory assessment context influenced by the location of infelicities in the essay (early or late)?

2.1 To what extent are raters' sequential, paragraph-by-paragraph evaluative impressions of an essay in a laboratory assessment context influenced by the rhetorical sophistication of the introductory paragraph of the essay?

2.2 To what extent are raters' sequential, paragraph-by-paragraph evaluative impressions of an essay in a laboratory assessment context influenced by the presence or location of organizational consistency (consistency, inconsistency at paragraph two, inconsistency at paragraph four) in the body of an essay relative to the organizational plan established in the first paragraph of that essay?

2.3 To what extent are raters' sequential, paragraph-by-paragraph evaluative impressions of an essay in a laboratory assessment context influenced by the density (early versus late) of error in the essay?

STUDY 3

3.0 Given a portfolio consisting of three writing samples, does the inclusion of a writing sample of higher or lower quality than the other two samples affect the ratings of the portfolio as a whole?

3.1 Does the position of the dissonant writing sample in relation to the other two samples influence the rating of the portfolio as a whole? That is, does either the primacy or recency of a dissonant quality writing sample within a portfolio affect the overall evaluation of that portfolio?

In the broadest of strokes, results supported the conclusion that raters of individual essays are indeed disproportionately influenced by first impressions. Textual features that come later in an essay generally cannot overcome quality perceptions already formed. In addition and as expected, raters were generally favorably impressed and more forgiving when they encountered

sophisticated introductions or when the organizational structure promised in the introductory paragraph was in fact delivered. While these first impression effects dominated evaluations of individual essays, portfolio evaluations as conducted in this project seemed relatively immune.

Summary of Findings, Study 1

The most important finding in Study 1 was the significant decrement in mean ratings assigned to essays with high error rate in the early part of the essay, as opposed to essays with exactly the same types and frequencies of error, but delayed until the latter part of the essay. Logically, the quality of the two sets of essays ought to have been equivalent; after all, they both contained the same mechanical errors. The finding instead of a significant disadvantage for early-appearing errors, unredeemed by late-appearing mechanical correctness, indicated that raters operating under conditions of duress typical of large-scale high-stakes assessment rating sessions formed initial impressions of quality that resisted subsequent information to the contrary.

Although the results of this study did indicate a statistically significant tendency for error location to affect average judgments about overall composition quality, that tendency was not so potent that it affected high-stakes pass/fail outcomes in the minimum competency test of writing ability.

Summary of Findings, Study 2

Study 2, the most complex of the three studies undertaken in this dissertation, focused on raters' reactions to a number of permutations of textual features encountered as they read to assess the quality of essays that had been manipulated for three variables of interest: sophistication of introductory strategy (two levels), organizational consistency (three levels), and location of error density (two levels). Study 2 examined the influence of this set of textual

features as raters encountered them in assessing the quality of essays, paragraph by paragraph, throughout a set of experimental essays.

The analysis of Study 2 focused on two patterns of results. First, the analysis examined the unfolding, paragraph-by paragraph evaluative impressions of raters as those raters encountered the various combinations of infelicities instantiated in essays at strategic locations (early or late). Second, whole essay ratings (that is, the ratings assigned after the final paragraph of each essay) within the "simple effects" combinations of factors were compared. The results of these two sets of comparisons are presented in turn below.

Comparisons of Paragraph-by-Paragraph Impressions. This analysis focused on comparisons across real time as raters read each paragraph in succession. Some of the essays started out with basic, blunt introductions, and others with more sophisticated ones ("funnel" introductions). In each case certain objective changes were introduced at various points: a low rate of error switched to a higher rate or vice versa. In some cases violations of organizational plan occurred immediately after the introductory paragraph; in other cases those violations occurred further along in the essay; in some instances organizational consistency was maintained throughout the essay.

Despite these manipulated changes in textual factors, in seven of the twelve treatment combinations raters never moved off their initial paragraph ratings as they proceeded through the remaining four paragraphs. That is, the composition ratings they delivered at paragraph 1 were the same as they registered after presumably reading the entire essay—and at every point in between. It is always dangerous to draw inferences from a failure to reject a null hypothesis. Nonetheless, that lack of responsiveness to objective changes in essay features cannot automatically be attributed to low statistical power. Observed power was .971, so it should have

been possible to detect significant differences had they existed. Furthermore, a host of other contrasts within this ANOVA effect were statistically significant.

Rather, a very plausible explanation is that in a slight majority of scenarios, once raters formed an initial judgment of composition quality at paragraph 1, they simply were unmoved by (or perhaps did not attend to) any further information that might reasonably have caused them to reassess their positions. Even midway through the essays, when the error rate increased in one set of papers or decreased in the other, the relative evaluations of the papers did not flip-flop in a corresponding fashion. It was as if the raters were unaffected by changes in error frequency that presented after the midpoint of the essays.

In five of the twelve treatment combinations, however, this scenario did not play out quite so neatly. When there were no violations of organizational expectations and the introductory strategy was relatively sophisticated ("funnel"), readers did appropriately register changes in mechanical correctness. That is, in this rarefied condition (organizational consistency and strong introduction), when error rate started high and then decreased, composition ratings did improve accordingly. Conversely, when error rate started low and then increased in later paragraphs, composition ratings did decline, as would be expected of an attentive rater.

When the introductory strategy was a simple statement of thesis and the first paragraphs had a low density of error, raters apparently did discern and react negatively to an organizational violation at paragraph 2. Unclouded by mechanical errors, therefore, raters did respond as expected when they encountered a violation of organizational expectation.

One statistically significant set of contrasts of the twelve was aberrant. When the introductory strategy was simple and the first paragraphs contained a high density of error, scores did increase when the error rate improved in the final paragraphs of the essays. Curiously,

this was the case only when an organizational inconsistency was introduced at approximately the same point as the improvement in error rate. This latter pattern defies clear explanation.

Comparisons of Whole Essay Ratings. In addition to examining the unfolding evaluations of essays by comparing paragraph ratings within each of the treatment combinations, as above, the final paragraph ratings—equivalent to holistic essay ratings--were compared within "simple effects." Eighteen such contrasts were calculated. Nine proved statistically significant.

Five of these simple-effects contrasts involved comparisons across levels of organizational inconsistency. Essays that began with simple introductions, had error in early rather than in later sections, and maintained organizational consistency were rated higher than the otherwise similar essays that had subsequent organizational violations early in the essay (paragraph 2). Essays that had simple introductions, had error appearing in later rather than in earlier sections, and were organizationally consistent were rated higher than otherwise similar essays that manifested organizational inconsistency later in the essay (paragraph 4). Essays that began with simple introductions, had error in early rather than in later sections, and manifested organizational inconsistency late in the essay (paragraph 4) were rated higher than otherwise similar essays that manifested organizational inconsistency relatively earlier (paragraph 2). Essays that began with sophisticated introductions, had error appearing in later sections rather than in earlier sections, and manifested organizational inconsistency late in the essay (paragraph 4) were rated higher than the otherwise similar essays with inconsistency manifested earlier (paragraph 2).

The pattern thus far explicated indicated that the larger the portion of the essay read without encountering organizational inconsistency, the higher it was rated—all other factors held equal. That conclusion seems quite reasonable. However, this pattern was reversed in one of the

final paragraph comparisons. When essays began with the more sophisticated funnel introduction, had error appearing in later segments rather than in earlier segments, and manifested no organizational inconsistency, they were rated more poorly than otherwise similar essays which manifested organizational inconsistency later in the essay (paragraph 4). This latter contrast defies principled explanation. One would have expected the organizational consistency to have engendered higher, not lower, overall impressions.

One more pair of final essay comparisons revealed the advantage of sophisticated introductory strategies on ratings. When organizational inconsistency occurred early in the essay and error also occurred earlier rather than later in the essay, and a sophisticated funnel introduction was used, the essay was rated higher than the otherwise similar essays with the less sophisticated introduction. This latter essay would be presumed to have the greatest number disadvantageous features, and so it is no surprise that it was rated so poorly. Similarly, when organizational inconsistency occurred later in the essay, error density was greater in the later than in the earlier segments, and a sophisticated introduction was used, the essay was rated higher than otherwise similar essays with simple introductions.

Two additional contrasts on final ratings revealed the advantage of late-occurring error density. When an essay had a simple introduction, organizational inconsistency early rather than late, and error predominant in later rather than earlier sections, the essay received a higher rating than otherwise similar essays in which dense concentration of error occurred early rather than late. And when an essay had a sophisticated introduction, organizational inconsistency occurred later rather than earlier in the essay, and error occurred later rather than earlier, the essay was rated higher than otherwise similar essays in which error appeared earlier.

Thus, with the exception of one anomalous comparison, these contrasts between whole essay ratings indicated an advantage for organizational consistency, or at least late onset of organizational inconsistency, as compared with early organizational inconsistency. The findings also indicated an advantage for later-occurring error density over essays that had dense errors toward the beginning. Finally, this set of comparisons indicated an advantage for essays which early on are relatively clean mechanically (even though they may deteriorate in their closings), as compared to essays dense with errors at the beginning but having few errors in the latter half.

Summary of Findings, Study 3

The purpose of Study 3 was to determine the degree to which raters' holistic evaluations of portfolios were disproportionately influenced by dissonant quality extemporaneous essays embedded within them, and whether the impression rendered by first-appearing dissonant essays unduly influenced overall judgments of those portfolios. Previous research bearing on this subject that utilized rater protocol analysis gave reason to suppose that raters' holistic perceptions would be largely influenced by the quality of the first elements they encountered in those portfolios (Hamp-Lyons & Condon, 1993). The findings of the Study 3, however, did not support that conclusion. To the contrary, raters in this study seemed to adhere to the instructions given them to give equal weight to all components of the portfolios. Where dissonant quality extemporaneous essays appeared, they were not given disproportionate weight. Moreover, Study 3 failed to reveal any differential impact for the position within the portfolio in which a higher or lower quality essay appeared.

Situating the Research Within the Field of Writing Assessment

This research adds to the knowledge base about the reading process of raters engaged in formal writing assessment. This process has been described most notably by Freedman & Calfee

(1983). In particular, these dissertation studies were designed to add to the small body of research investigating the nature and integrity of the holistic reading process as a dynamic phenomenon subject to many interactions of rater, textual features, and assessment context. The three studies investigated questions about the judgmental processes of holistic raters as they evaluated individual essays or portfolios of student writing in real time. In particular, the question of how early text samples function as contextual or priming mechanisms to affect overall rater judgments of essays or portfolios was addressed.

The research studies conducted for this dissertation provide some evidence that, in large-scale writing assessments, holistic raters are more influenced by what appears early in an essay than by what is written in the second half of essays. The findings of Study 1, in particular, support an "early impression" hypothesis that some evaluative impressions formulated by raters early in the process of reading a writing sample exercise a disproportionate influence on the final rating they assign to that writing sample. Likewise, in Study 2, there was some evidence that initial impressions carried the day when it came to overall ratings.

Further dramatic evidence for the overriding impact of early impressions derived from paragraph-by-paragraph rating data for those most infelicitous of essays, those leading off with simple introductions, dense concentration of errors, and organizational inconsistency appearing in the first body paragraph. Even when these essays improved in the second half with a substantially diminished error rate and the resumption of the organizational plan presented in the opening paragraph, raters as a group did not raise their scores significantly.

However, it was not just in cases of early "triple jeopardy" that raters' initial impressions continued unchanged throughout essays. In spite of the balanced manipulation of textual features in the first and second halves of essays, raters never shifted from their first-paragraph

quality ratings in a slight majority of treatment conditions in the study, specifically, seven out of twelve of them. In a number of these cases, raters' lack of responsiveness to objective changes in texts seemed to be associated with the level of error density established early in the essays. In several treatment conditions, essays that began with low error rates were rated at every paragraph more favorably than papers which began with a high concentration of errors in the first half. In these instances, raters did not respond to diminished mechanical control manifested after the midpoint of essays.

In some treatment combinations, however, initial impressions did not unilaterally carry the day. In these cases, raters' scores did reflect discernment of improvement or decline from the first half of essays to the second half. When essays began strongly with more sophisticated introductions and lower concentrations of error, and maintained organizational consistency to the end, raters were able to react appropriately when error intensified in the second half. Scores declined in spite of the strong first half and the maintenance of organizational consistency in the second half. Conversely, composition ratings improved as errors diminished in the second half of essays after those essays began with strong introductions and organizational consistency. This particular set of findings argues against the unvarnished potency attributed to error rate in some earlier composition research (e.g., Rafoth & Rubin, 1984). Based on the present findings, it appears that essays with a variety of sources of disrepair can interfere even with raters' perception of mechanical error.

The findings discussed thus far are relevant to issues of validity regarding holistic writing assessment. The findings from Studies 1 and 2 that provide support for the early impression hypothesis indicate a threat to validity. The tendency of some raters to be disproportionately influenced by the earlier sections of texts, as opposed to the later sections, means their holistic

ratings of the overall quality of a writing sample may undervalue what happens in the essay subsequent to the sections that influenced those early negative impressions. Thus, essays to which raters respond favorably on the basis of early appearing text may be rated favorably overall, even if the quality of the essay deteriorates in later sections. Conversely, essays which begin badly but recover in later sections may be undervalued by the overall holistic rating.

Results of Study 3, in contrast, offer some support to advocates of portfolio assessment as an antidote to the deficiencies of single-essay holistic rating. Study 3 failed to produce evidence that holistic raters of writing portfolios gave undue influence to essays placed early in the portfolios, contradicting the findings of Hamp-Lyons & Condon (1993). Not only do these findings fail to provide support for the early impression hypothesis as it bears on portfolio assessment, they also contradict previous research indicating that raters are unable to equitably blend the quality of portfolio component essays into a single portfolio impression (Herman, Gearhart, & Baker, 1993). In the present study of portfolio assessment, in fact, raters appeared to give equal weight to all the essays comprising each portfolio, just as they were instructed to do during training.

It should be noted that the type of portfolio assessment studied here was characterized by writing samples of uniform genre (expository essay) and comparable length. Under those standardized conditions, raters were able to adhere to the scoring guidelines they were instructed to follow. This finding is in keeping with the research of Nystrand, Cohen and Dowling (1993), which produced findings suggesting that a greater degree of standardization in the elements making up portfolios may improve the reliability of portfolio assessments. On the other hand, many of the proponents of portfolio assessment have in mind an operation considerably less standardized than what was implemented here (for example, Elbow & Belanoff, 1991; Huot,

2002; Moss, 1994; Wiggins, 1994; Yancey, 1992). Findings regarding rater behavior in the present study should not be loosely and uncritically generalized to such informal and diverse portfolio assessment systems.

Although a considerable body of previous research had examined mechanical error as a powerful source of influence on evaluators, scant research had empirically tested common pedagogical injunctions to utilize sophisticated introductory strategies. Although Study 2 did not unambiguously support a claim of higher ratings for essays which began with the more sophisticated of the two types of introductions studied here, there is evidence that beginning with a "funnel" introduction can at least mitigate the potential negative fallout of other infelicities.

Prior to the present study only a small body of empirical studies (Connor, 1987; Ferris, 1994; Hake, 1986) examined the impact of explicit organizational cues on judged composition quality. Study 2 extended the investigation of raters' reactions to organization as a textual feature by manipulating organizational consistency in experimental essays. It built upon a theoretical thread in Frank Smith's (1994) description of the comprehension processes of skilled readers. The more successfully writers enable readers to anticipate formal structures, Smith theorized, the more successful the reader's comprehension. Inversely, the more discrepant the interaction between a reader's predictions and the formal structures of a text, the less successful the reader's understanding of the text. Results of Study 2 did indicate that organizational consistency interacts with other textual features to affect composition quality. Essays that maintained organizational consistency, or that delayed organizational inconsistency until late in the essay at paragraph 4, fared relatively well in terms of quality ratings.

One of the research projects conducted for this dissertation is significant for an important methodological contribution for the field of writing assessment. Study 2 met two

methodological challenges commonly associated with investigations of the influences on raters' decisions in various assessment contexts. First, the online reporting methodology (private, minimally intrusive, paragraph-by-paragraph reports of impressions of writing quality) of Study 2 avoided the intrusiveness of protocol analyses that are employed—sometimes dysfunctionally (see Smagorinsky, 1989, 1998)—to try to capture the real-time development of judgments about writing quality. Second, this online methodology allowed easier collection of more data than is commonly feasible with studies relying on labor- and time-intensive protocol analysis (e.g., Martin, 1987 and Vaughan, 1991). The limited sample sizes (of both raters and the number of essays examined) characteristic of such studies does not contribute to their persuasiveness (Wolfe, 1997). The online reporting methodology used in Study 3 therefore offers a significant advantage for future studies that seek to investigate the dynamic nature of evaluative reading processes.

One additional way in which this research has responded to a challenge in the field of writing assessment has to do with the naturalistic assessment context used in Study 1. Huot (1993) has urged more studies of raters operating in realistic rather than laboratory settings. The placement in Study 1 of essays manipulated to allow experimental examination of the behavior of raters in an actual statewide, high-stakes assessment eliminated two of the limitations of most writing assessment studies: small sample size and the artificiality of rater psychology in laboratory contexts. Because the assessment context was naturalistic and raters, who were unaware of the inclusion of manipulated essays in the assessment, were actually engaged in a large-scale, high-stakes assessment, the psychology of raters involved was authentic. Because of the high number of rater participants (N = 88), the statistical power of the analysis was high. And because the rating decisions were of genuine high-stakes importance in pass/fail decisions

for students, the findings (evidence which supports the early impression hypothesis) of the study are unquestionably meaningful.

Implications for Practice

Rater training is emphasized in every guideline about implementing writing performance assessment (see, for example, White, 1998). The evidence from Studies 1 and 2 highlights the need for training procedures which will inoculate raters against the tendency to become mired in their early impressions of an essay's quality. How best to go about such inoculation is at present unknown, but it is reasonable to expect that revealing to raters their own discrepant judgments based solely on the location of mechanical errors could at least caution them to withhold early judgments and attend to entire essays. Unlike professionals in some fields of expertise, expert essay readers do not formulate early diagnoses which then require dramatic disconfirmation to change. Rather, they keep their minds open throughout the process of reading an essay (Huot, 1993; Wolfe, 1997).

The evidence from Study 3, which failed to support the early impression hypothesis as it applies to portfolio assessment, reinforces the value of explicit instructions and training raters to give appropriate weighting to all writing samples contained in a portfolio that is to be rated holistically. In the latter instance, raters were specifically instructed to consider all the essays in each portfolio equally in formulating their holistic rating of the quality of that portfolio. In contrast, neither the written instructions for scoring essays holistically (see Appendix A) used in Studies 1 and 2, nor the oral instructions given to raters during training sessions for those studies, encouraged raters to be sure to give equal weighting to all parts of each essay they read. On the other hand, some earlier research on composition ratings found little impact of explicit directions on rater perceptions (Breland & Jones, 1984; Harris, 1977; Rafoth & Rubin, 1984).

Proponents of portfolio assessment may find some considerable confirmation in the results of this study. Portfolio rating did appear to facilitate a clear perception and balancing of the component essays within the portfolios. Perhaps one advantage of portfolio assessment vis à vis construct validity lies in the fact that reading an entire portfolio is a slower and necessarily more iterative process than gaining an impression from a single essay. No doubt that additional time and reflection impose real costs that directors of assessment programs need to factor into their plans and budgets. Furthermore, it must be emphasized once again that the version of portfolio assessment implemented in the present study was of the most controlled and standardized nature, a far cry from the use of portfolios sometimes advocated for student self development (e.g., encouraging students to select their own exhibits, including early drafts, and binding the whole with self-reflective essays; see Huot, 2002; White, 1998; Yancey, 1992). The model of portfolio assessment utilized here is relevant only to program assessment or selection/certification decisions, in which writers are anonymous and are responding to uniform elicitation prompts.

Limitations

For all three studies conducted for this dissertation, generalizability of the writing samples used as the basis of measurements of writing ability is open to question. Whether single-sitting extemporaneous writing samples or multiple samples composed over time, the writing samples included in these studies might be challenged as to whether they are representative of other samples belonging to other genres or different writing conditions. Further, the generalizability of raters' behaviors in the studies might be questioned; whether other raters, even those with similar training and experience, would behave similarly is open to question. In defense against such charges of lack of generalizability, the methods employed

here should be regarded as conforming quite closely only to a particular statewide postsecondary writing assessment program. Stimulus essays were elicited within the context of the Georgia RTPET, raters were experienced Georgia RTPET evaluators, and the scoring rubrics used were identical to those used in the Georgia RTPET. Beyond that isomorphism, these studies can guarantee no further generalizability.

In addition, the motivation of raters in Studies 2 and 3 must be considered as a limitation. Because these studies are laboratory studies, the psychology of raters involved can not be the same as if they were actually engaged in authentic, high-stakes assessments. On the other hand, the motivation of raters in Study 1 was quite authentic. Raters in that study could not distinguish between a high-stakes decision and one that contributed only to dissertation research.

One of the limitations of the research methodology used in Study 2, though it represents a substantial advantage over think-aloud protocol analysis, pertains to the artificiality of asking raters to report their developing evaluative impressions of an essay at the end of each paragraph. Relative to think-aloud protocol elicitation, the computer assisted method used in Study 2 was nonintrusive, but its residual level of intrusiveness may to some degree have interfered with the natural formation of raters' judgments. Huot (1993) and Wolfe (1997) have observed that experienced and proficient raters tend to suspend judgment until they have read the entire essay. Because of the methodology used in this laboratory study, the raters in this study may have been more attentive than usual to their incremental impressions of the texts as they proceeded paragraph by paragraph through essays. This limitation was necessary in this study, however, in order to track the sequential reactions of raters to essays in which textual features have been systematically manipulated.

Future Research

It seems unlikely that writing assessment research will return to vigorous investigation of the influence of various textual features on raters, the kinds of questions that were commonly explored in the 1960s, `70s, and `80s. On the other hand, high-stakes, large-scale assessments show few signs of going away in the 21st Century, despite the higher profile of portfolio assessment in the classroom and in local institutional contexts (e.g., course placement) beyond the classroom. The terrain explored in this set of research projects, that is, the dynamic nature of the reading processes involved in writing assessment, presents fresh opportunities for research. The moment-by-moment flux of impressions that register in the minds of raters as they read to evaluate writing samples remains one of the more under-explored realms of large-scale writing assessment.

One area which deserves investigation is the question whether raters can, in fact, be trained to avoid the early impression syndrome, evidence for which was found in Studies 1 and 2. It is one thing to call attention to this unconscious pattern of; it is quite another to devise and implement training regimens to enable raters to achieve this goal in practice. Research projects that target this question might pay particular attention to the effects of fatigue on raters' behavior over an entire assessment period. In particular, such projects should examine whether raters are more prone to the early impression syndrome later in an assessment than they are at earlier stages of the assessment.

Another of the more compelling areas for continued investigation involves closer examination of the variable responses of raters to extemporaneous writing samples and to writing samples that have had the benefit of cycles of revision and feedback. Given current pedagogical emphases on collaborative writing and on project-based pedagogy, some writing instructors may

have relatively little opportunity to teach or to evaluate the kind of independent, timed, and supervised writing tasks presumed in formal assessments of writing ability. In the present study, raters were not disproportionately influenced by an essay designated as "extemporaneous." Future studies might systematically vary the degree to which a writing sample may be taken to reflect a student's independent efforts. Does a paper explicitly identified as a group project carry as much weight in judging writing proficiency as one identified as authored by a single student but with the benefit of peer feedback?

A related line of research might query various assessment stakeholders regarding their expectations and the credibility which they accord to writing samples elicited with varying degrees of student autonomy. For example, do school administrators seeking to evaluate instructional programs feel uncomfortable when no supervised extemporaneous essay appears in student portfolios? Do perhaps employers, who increasingly desire professionals who can work collaboratively, prefer to see evidence of proficiency in joint writing assignments? Do the demands for raters to maintain fidelity to formal rating rubrics—central to the three studies in this dissertation—vary according to the uses to which stakeholders are putting evaluation data?

The methodological protocols devised in conjunction with this study—e.g., embedding experimental essays in high stakes ratings or using computer assisted means to capture unfolding perceptions of an essay—hold considerable promise for future studies which might manipulate a different set of textual features. In particular, they may be applied to studies which examine the impact of text markers of socio-cultural identity, which in turn may be associated with rater social biases of various kinds. For example, some authors have asserted that certain textual features—politeness markers or expletives—elicit rather clear gender-typed images of student writers (Haswell & Haswell, 1996). Other studies have examined the effects on essay evaluation

of students' ethnic identities that may be triggered by textual features (see discussion in Rubin, 1995). Important questions of equity revolve around the processes by which raters make gender or ethnic attributions of student writers and whether text elements more objectively associated with writing quality can overcome any biased expectations that raters may evolve early in an essay reading. The tools described in the present study are eminently suitable for such research.

REFERENCES

Anson, C.M., & Brown, R.L. (1991). Large-scale portfolio assessment. In P. Belanoff & M. Dickson (Eds.), *Portfolios: Process and product* (pp. 248-269). Portsmouth, NH: Boynton/Cook.

Applebee, A.N. (1991). Environments for language teaching and learning: Contemporary issues and future directions. In J. Flood, J.M. Jenson, D. Lapp, & J.R. Squire (Eds.), *Handbook of research on teaching the English language arts* (pp. 549-556). New York: International Reading Association & National Council of Teachers of English.

Baker, S. (1998). *The practical stylist* (8th ed.). New York: Longman.

Barker, L.L., Kibler, R.J., & Hunter, E.C. (1968). An empirical study of overlap rating effects. *The Speech Teacher, 17*, 160-166.

Barritt, L., Stock, P.L., & Clark, F. (1986). Researching practice: Evaluating practice essays. *College Composition and Communication, 37*, 315-327.

Braddock, R., Lloyd-Jones, R., & Schoer, L. (1963). *Research in written composition*. Urbana, IL: National Council of Teachers of English.

Breland, H., & Jones, R.J. (1984). Perceptions of writing skills. *Written Communication, 1*, 101-109.

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing.* Logan: Utah State University Press.

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: Heinemann.

Brossell, G. (1996). Writing assessment in Florida: A reminiscence. In E.M. White, W.D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 25-32). New York: Modern Language Association.

Camp, R. (1993). Changing the model for the direct assessment of writing. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 45-78). Creskill, NJ: Hampton Press.

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

Clark, I. L. (1993). Portfolio evaluation, collaboration, and writing centers. *College Composition and Communication, 44*, 515-524.

Connor, U. (1987). Argumentative patterns in student essays: Cross-cultural differences. In U. Connor & R.B. Kaplan (Eds.), *Writing across languages: analysis of L2 text* (pp. 57-71). Reading, MA: Addison-Wesley.

Cooper, C.R. (1977). Competency testing: Issues and overview. In C.R. Cooper (Ed.), *The nature and measurement of competency in English* (pp. 1-20). Urbana, IL: National Council of Teachers of English.

Crowhurst, M. (1980). Syntactic complexity and teachers' ratings of narrations and arguments. *Research in the Teaching of English, 14*, 223-231.

Diederich, P.B. (1974). *Measuring growth in English*. Urbana, IL: National Council of Teachers of English.

Diederich, P.B., French, J.W., & Carlton, S.T. (1961). *Factors in judgments of writing ability.* Princeton, NJ: ETS. ERIC reprint document # ED002172.

Eggers, P. (1998). *Process and practice: A guide for developing writers* (4th ed.). New York: Longman.

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

Elbow, P. (1994). Will the virtues of portfolios blind us to their potential dangers? In L. Black, D. Daiker, J. Sommers, & G. Stygall (Eds.), New directions in portfolio assessment (pp. 40-55). Portsmouth, NH: Boynton/Cook.

Elbow, P. (1996). Writing assessment: Do it better, do it less. In E. M. White, W. D. Lutz, and S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 120-134). New York: Modern Language Association.

Elbow, P., & Belanoff, P. (1991). State University of NY at Stony Brook portfolio-based assessment program. In P. Belanoff and M. Dickson (Eds.), *Portfolios: process and product* (pp. 3-16). Portsmouth, NH: Boynton/Cook.

Farr, M. (1996). Response: Awareness of diversity. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 241-244). New York: Modern Language Association.

Farr, R., & Beck, M. (1991). Evaluating language development: Formal methods of evaluation. In J. Flood, J.M. Jenson, D. Lapp, & J.R. Squire (Eds.), *Handbook of research on teaching the English language arts* (pp. 489-501). New York: International Reading Association & National Council of Teachers of English.

Fawcett, S. (2004). *Evergreen: A guide to writing with readings* (7th ed.). Boston: Houghton Mifflin.

Ferris, D.R. (1994). Rhetorical strategies in student persuasive writing: Differences between native and non-native speakers. *Research in the Teaching of English, 28*, 45-65.

Fish, S.E. (1980a). *Is there a text in this class? The importance of interpretive communities.* Cambridge, MA: Harvard University Press.

Fish, S.E. (1980b). Literature in the reader: Affective stylistics. In J.P. Tompkins (Ed.), *Reader-response criticism* (pp. 70-100). Baltimore: Johns Hopkins University Press.

Flower, L. (1988). Constructing a purpose in writing and reading. *College Composition and Communication, 50*, 528-550.

Flower, L. (1989). Cognition, context, and theory building. *College Composition and Communication, 40*, 282-311.

Frederiksen, C.H. (1986). Cognitive models and discourse analysis. In C.R. Cooper & S. Greenbaum (Eds.), *Studying writing: Linguistic approaches* (pp. 227-267). Beverly Hills, CA: Sage.

Freedman, S.W. (1979). How characteristics of student essays influence teachers' evaluation. *Journal of Educational Psychology, 71*, 328-338.

Freedman, S.W. (1981). Influences on the evaluation of expository essays: Beyond the text. *Research in the Teaching of English, 15*, 245-255.

Freedman, S.W. (1993). Linking large-scale testing and classroom portfolio assessments of student writing. *Educational assessment, 1* (1), 27-52.

Freedman, S.W., & Calfee, R.C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor, & S.A. Walmsley (Eds.), *Research in writing: Principles and methods* (pp. 75-98). New York: Longman.

Gearhart, M., Herman, J.L., Novak, J.R., & Wolf, S.A. (1995). Toward the instructional utility of large-scale writing assessment: Validation of a new narrative rubric. *Assessing Writing, 2* (2), 207-242.

Gere, A.R. (1980). Written composition: Toward a theory of evaluation. *College English, 42*, 44-48.

Glaser, R., & Silver, E. (1994). Assessment, testing, and instruction: Retrospect and prospect. In L. Darling-Hammond (Ed.), *Review of research in education, 20*, 393-421.

Godshalk, F.I., Swineford, F., & Coffman, W.E. (1966). *The measurement of writing ability.* (Research Monograph No. 6). New York: CEEB.

Grobe, C. (1981). Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. *Research in the Teaching of English, 15*, 75-85.

Hake, R. (1986). How do we judge what they write? In K.L. Greenberg, H.S. Wiener, & R.A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 153-167). New York: Longman.

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

Harris, W.H. (1977). Teacher response to student writing: A study of the response patterns of high school English teachers to determine the basis for teacher judgment of student writing. *Research in the Teaching of English, 11*, 175-185.

Haswell, R.H., & Haswell, J.T. (1996). Gender bias and critique of student writing. *Assessing Writing, 3* (1), 31-83.

Haswell, R.H. (Ed.) (2001). *Beyond outcomes: Assessment and instruction within a university writing program*. Westport, CT: Ablex.

Herman, J.L., Gearhart, M., & Baker, E.L. (1993). Assessing writing portfolios: Issues in the validity and meaning of scores. *Educational Assessment, 1*(3), 201-224.

Hillocks, G., Jr. (2002). *The testing trap: How state writing assessments control learning*. New York: Teachers College Press.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*(2), 237-263.

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. (1996). Toward a new theory of writing assessment. *College Composition and Communication, 47* (4), 549-566.

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

Johnston, P. (1989). Constructive evaluation and the improvement of teaching and learning. *Teachers College Record, 90*, 509-528.

Kamusikiri, S. (1996). African-American English and writing assessment: An Afrocentric approach. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 187-203). New York: Modern Language Association.

Langan, J. (2001). *College writing skills with readings* (5th ed.). New York: McGraw-Hill.

Larson, R.L. (1996). Portfolios in the assessment of writing: A political perspective. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 271-283). New York: Modern Language Association.

Lloyd-Jones, R. (1977). Primary trait scoring. In C. Cooper & L. Odell (Eds.), *Evaluating writing*. Buffalo, NY: National Council of Teachers of English.

Lucas, C. (1992). Introduction: Writing portfolios—Changes and challenges. In K.B. Yancey (Ed.), *Portfolios in the writing classroom: An introduction* (pp. 1-11). Urbana, IL: National Council of Teachers of English.

Lunsford, A., & Connors, R. (1995). *The St. Martin's handbook* (3rd ed.). New York: St. Martin's Press.

Mabry, L. (1999, May). Writing to the rubric: Lingering effects of traditional standardized testing on direct writing assessment. *Phi Delta Kappan*, 673-679.

Madaus, G.F., & O'Dwyer, L.M. (1999, May). A short history of performance assessment: Lessons learned. *Phi Delta Kappan*, 688-695.

Markham, L. R. (1976). Influences of handwriting quality on teacher evaluation of written work. *American Educational Research Journal, 13*, 277-283.

Martin, W. (1987). *A study of reader process in the evaluation of English placement essays.* Unpublished doctoral dissertation, University of Louisville.

McCulley, G.A. (1985). Writing quality, coherence, and cohesion. *Research in the Teaching of English, 19*, 269-282.

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

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23* (2), 13-23.

Moss, P.A. (1994a). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12.

Moss, P.A. (1994b). Validity in high stakes writing assessment. *Assessing Writing, 1*, 109-128.

Myers, M., & Pearson, P.D. (1996). Performance assessment and the literacy unit of the new standards project. *Assessing Writing, 3* (1), 5-29.

Nielsen, L., & Piche, G. (1981). The influence of headed nominal complexity and lexical choice on teachers' evaluation of writing. *Research in the Teaching of English, 15*, 65-74.

Neuner, J.L. (1987). Cohesive ties and chains in good and poor freshmen essays. *Research in the Teaching of English, 21*, 91-105.

Nold, E.W., & Freedman, S.W. (1977). An analysis of readers' responses to student essays. *Research in the Teaching of English, 11*, 164-174.

Nystrand, M., Cohen, A.S., & Dowling, N.M. (1993). Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment, 1* (1), 53-70.

Pula, J.J., & Huot, B.A. (1993). Model of influences on holistic raters. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 237-265). Cresskill, NJ: Hampton Press.

Rafoth, B.A., & Rubin, D.L. (1984). The impact of content and mechanics on judgments of writing quality. *Written Communication, 1*, 446-458.

Ramirez, A. (1999, November). Assessment-driven reform: The emperor still has no clothes. *Phi Delta Kappan*, 204-208.

Raymond, J.C. (1982). What we don't know about the evaluation of writing. *College Composition and Communication, 33*, 399-403.

Reither, J.A., & Hunt, R. (1993). Beyond portfolios: Scenes for dialogic reading and writing. In L. Black, D. Daiker, J. Sommers, & G. Stygall (Eds.), *New directions in portfolio assessment* (168-182). Portsmouth, NH: Boynton/Cook.

Resnick, L.B., & Resnick, D.P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B.R. Gifford, & M.C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37-75). Boston: Kluwer Academic Publishers.

Rosenblatt, L.M. (1938, 1976). *Literature as exploration* (3rd ed.). New York: Modern Language Association.

Rosenblatt, L.M. (1978). *The reader, the text, the poem.* Carbondale, IL: Southern Illinois University Press.

Rubin, D.L., & Mead, N.A. (1984). *Large scale assessment of oral communication skills: Kindergarten through grade 12. Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills.*

Rubin, D.L. (Ed.) (1995). *Composing social identity in written language.* Hillsdale, NJ: Lawrence Earlbaum Associates.

Ruddell, R.B., & Unrau, N.J. (1994). Reading as a meaning-construction process: The reader, the text, and the teacher. In R. Ruddell, M. Ruddell, & H. Singer (Eds.), *Theoretical models and processes of reading* (4th ed.), (pp. 996-1056). Newark, DE: International Reading Association.

Scannell, D.P., & Marshall, J.C. (1966). The effect of selected composition errors on grades assigned to essay examinations. *American Educational Research Journal, 3*, 125-130.

Shaughnessy, M. (1977). *Errors and expectations*. New York: Oxford University Press.

Smagorinsky, P. (1989). The reliability and validity of protocol analysis. *Written Communication, 6*, 463-479.

Smagorinsky, P. (1998). Thinking and speech and protocol analysis. *Mind, Culture, and Activity, 5* (3), 157-177.

Smith, F. ((1997). Reading without nonsense (3rd ed.). New York: Teachers College Press.

Smith, F. (1994). *Understanding reading: A psycholinguistic analysis of reading and learning to read* (5th ed.). New York: Holt, Rinehart & Winston.

Smith, F. (2003). Unnatural acts, unnatural practices: Flaws and fallacies in "scientific" reading instruction. Portsmouth, NH: Heinemann.

Smith, W.L. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement program technique. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp.142-205). Creskill, NJ: Hampton Press.

Stewart, M.R., & Grobe, C.H. (1979). Syntactic maturity, mechanics, and vocabulary and teachers' quality ratings. *Research in the Teaching of English, 13*, 207-215.

Stotsky, S. (1979). Teaching the vocabulary of academic discourse. *Journal of Basic Writing, 2*, 15-39.

Stotsky, S. (1981). The vocabulary of essay writing: Can it be taught? *College Composition and Communication, 32*, 317-326.

Taylor, C. (1994). Assessment for measurement or standards: The peril and promise of large-scale assessment reform. *American Educational Research Journal, 31*, 231-262.

Tierney, R.J., Carter, M.A., & Desai, L.E. (1991*). Portfolio assessment in the reading-writing classroom.* Norwood, MA: Christopher-Gordon.

Tierney, R.J., & Mosenthal, J.H. (1983). Cohesion and textual coherence. *Research in the Teaching of English, 17*, 215-229.

Underwood, T., & Murphy, S. (1998). Interrater reliability in a California middle school English/language arts portfolio assessment program. *Assessing Writing, 5* (2), 201-230.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language in academic contexts* (pp. 111-126). Norwood, NJ: Ablex.

Veal, L.R. (1974). *Syntactic measures and rated quality in the writing of young children.* (Studies in Language Education, Report # 8). Athens, GA: University of Georgia, (ERIC Reproduction Service No. 090 55).

Veal, L.R., & Hudson, S.A. (1983). Direct and indirect measures for large-scale evaluation of writing. *Research in the Teaching of English, 17*, 290-296.

White, E.M. (1993). Holistic scoring: Past triumphs, future challenges. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 79-108). Cresskill, NJ: Hampton Press.

White, E.M. (1996a). Power and agenda setting in writing assessment. In E.M. White, W.D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 9-24). New York: Modern Language Association.

White, E.M. (1996b). Response: Assessment as a site of contention. In E.M. White, W.D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 301-304). New York: Modern Language Association.

White, E.M. (1996c). Writing assessment in the 21st century: Will writing teachers play a role? In L.Z. Bloom, D.A. Daiker, & E.M. White (Eds.), *Composition in the 21st century: Crisis and change* (pp. 101-111). Carbondale, Illinois: Southern Illinois University Press.

White, E.M. (1998). *Teaching and assessing writing* (2nd ed.). Portland, ME: Calendar Island.

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

Williamson, M.M. (1993). An introduction to holistic scoring: The social, historical and theoretical context for writing assessment. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 1-44). Cresskill, NJ: Hampton Press.

Witte, S.P., Daly, J.A., & Cherry, R.D. (1986). Syntactic complexity and writing quality. In D.A. McQuade (Ed.), *The territory of language* (pp. 150-164). Carbondale, IL: Southern Illinois University Press.

Witte, S.P., & Faigley, L. (1981). Coherence, cohesion, and writing quality. *College Composition and Communication, 32*, 189-204.

Wolcott, W., with Legg, S.M. (1998). *An overview of writing assessment.* Urbana, IL: National Council of Teachers of English.

Wolfe, E.W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4* (1), 83-106.

Yancey, K.B. (1992). Portfolios in the writing classroom: A final reflection. In K.B.Yancey (Ed.), *Portfolios in the writing classroom: An introduction* (pp. 102-116). Urbana, IL: National Council of Teachers of English.

Yancey, K.B. (1999, February). Looking back as we look forward: Historicizing writing assessment. *College Composition and Communication, 50* (3), 483-503.

APPENDICES

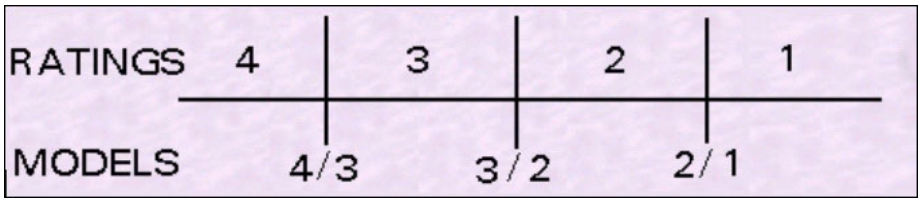
Appendix A

INSTRUCTIONS FOR SCORING REGENTS' TESTING PROGRAM ESSAYS

DESCRIPTION OF ESSAY SCORING PROCEDURE

Raters should read each essay quickly to gain a general impression of its quality in relation to the model essays and assign a rating based on that comparison. This approach, holistic rating, contrasts with the analytic grading commonly used in essay evaluation, but evidence indicates that holistic rating is much faster and produces more uniform results.

The essays are rated on a four-point scale in which "1" is the lowest score and "4" is the highest score. The model essays represent borderline cases; each essay to be rated must, by definition, fall above or below a model.



One model essay represents each dividing line. An essay better than the "2/1" model and worse than the "3/2" model would be rated "2." An essay worse than the "2/1" model becomes "1." An essay better than the "4/3" model becomes "4."

Note carefully that raters should compare the essays they read with the models . They should not rate in terms of their usual grading standards or some abstract standard. They should not associate the ratings with the traditional grades A , B , C , D , F .

The testing subcommittee of the University System Academic Committee on English attempts to choose models by using the following definitions of competency, although it realizes that these definitions are by no means exhaustive.

4 : The "4" essay has a clear central idea that relates directly to the assigned topic. The essay has a clear organizational plan. The major points are developed logically and are supported with concrete, specific evidence or details that arouse the reader's interest. The essay reveals the writer's ability to select effective, appropriate words and phrases; to write varied, sophisticated sentences; to make careful use of effective transitional devices; and to maintain a consistent, appropriate tone. The essay is essentially free from mechanical errors, it contains no serious grammatical errors, and the ideas are expressed freshly and vividly.

3 : The "3" essay has a clear central idea that relates directly to the assigned topic. It contains most of the qualities of good writing itemized above. The essay generally differs from a "4" in

that it shows definite competence, but lacks distinction. The examples and details are pertinent, but may not be particularly vivid or sharply observed; the word choice is generally accurate, but seldom -- if ever -- really felicitous. The writer adopts an appropriate, consistent tone. The essay may contain a few errors in grammar and mechanics.

2 : The "2" essay meets only the basic criteria, and those in a minimal way. The essay has a central idea related directly to the assigned topic and presented with sufficient clarity that the reader is aware of the writer's purpose. The organization is clear enough for the reader to perceive the writer's plan. The paragraphs coherently present some evidence or details to substantiate the points. The writer uses ordinary, everyday words accurately and idiomatically and generally avoids both the monotony created by series of choppy, simple sentences and the incoherence caused by long, tangled sentences. Although the essay may contain a few serious grammatical errors and several mechanical errors, they are not of sufficient severity or frequency to obscure the sense of what the writer is saying.

1 : The "1" essay has any one of the following problems to an extraordinary degree or it has several to a limited degree: it lacks a central idea; it lacks a clear organizational plan; it does not develop its points or develops them in a repetitious, incoherent, or illogical way; it does not relate directly to the assigned topic; it contains several serious grammatical errors; it contains numerous mechanical errors; ordinary, everyday words are used inaccurately and unidiomatically; it contains a limited vocabulary so that the words chosen frequently do not serve the writer's purpose; syntax is frequently rudimentary or tangled; or the essay is so brief that the rater cannot make an accurate judgement of the writer's ability.

2/1 MODEL 2/1

TOPIC: WHY WOULD YOU LIKE OR DISLIKE OWNING YOUR OWN BUSINESS?

Going out of Business Sale! Signs of this nature can be seen everywhere. Today opening up a business can be scary, because of the extensive risk, high cost, and extreme stress.

The chief reason I would not want to start my own business is the great risk of failure. Today statistics show that four out of every six businesses fail within the first year. Those are not very good odds for one just starting his or her own business.

The second reason not to start my own business is the high cost of starting a business. Businesses take a great deal of money to get started, and for that matter to keep running. The first thing one has to do is find a place to put the business. Lots are very expensive. Then a building has to be built, and merchandise to fill the building has to

be purchased.

Finally owning a business can be stressful. Being ones own boss can be stressful to her or him by the way of having to make all of the important decisions, or can cause stress at home. The stress at home can be very detrimental to the marriage, or even the family as a whole.

Concluding this owning a business is just one big headache. On the other hand some people are very successful, and they got that way by taking the risk of owning their own business. I personally don't think that owning a business is worth the risk, when working for someone else is a lot safer.

3/2 MODEL 3/2

TOPIC: DISCUSS THE INFLUENCE THAT ADVERTISING HAS HAD ON YOUR LIFE OR THE LIVES OF YOUR FRIENDS.

Advertising has a large influence on my life and the lives of my friends. Advertising has an influence on the cars we drive, the clothes we wear, and the food we eat.

Advertising influences the cars my friends and I drive. The television commercials paint an unrealistic picture of how good life is once you own their product. For example, one of the commercials for Volvo implies that a person doesn't have class unless he drives a Volvo. According to the Cadillac commercial, a car can not be elegant unless it is a Cadillac. Magazine ads are very similar to television ads. Magazine ads show beautiful women and handsome men gathered around an automobile, and imply that the reader can be like the people in the ad.

Advertising has an influence on those clothes we wear. Television and magazines show hair-thin models wearing different articles of clothing. The ads for Jordache or Calvin Klein are a good example of this fact. My friends and I sometimes feel that if the clothes look good, then they must be made good. We also hope the clothes look as good on us as they did on the models.

Finally, advertising influences our eating habits. There are ads for hamburgers, hotdogs, pizzas, beer, candies, cakes, and the list keeps going. Pizza Inn gives us more of the things we like. The people at Burger King treat us right. Everyone wants to be an Oscar Mayer hot dog. Michelob wants us to put a little weekend in our week. Of

course, relief is spelled Rolaids. With ads like these facing us every night who could resist?

In conclusion, I'd like to say that advertising influences the way everybody lives. The cars we drive, the clothes we wear, and the food we eat are all a result of advertising.

4/3 MODEL 4/3

TOPIC: DISCUSS WHY PEOPLE ARE FASCINATED BY AMUSEMENT PARKS SUCH AS DISNEY WORLD AND SIX FLAGS.

People of all ages, shapes, sizes, financial statuses, and interests pour, in vast numbers each year into such amusement parks as Disney World and Six Flags. Why the fascination with these places, even to the point of repetitive visits? Each individual has his own reason, but there are a few common to all. Here in a make-believe world can be found something for everyone.

On stepping from a sometimes harsh, ugly world through the gates of a "magic kingdom," one can do for a short while anything he desires. Vicarious living, with all the thrills and dangers of adventure in faraway places or daring escapades unavailable in everyday life, is here for the price of a ticket. There are wild rides: twisting, dipping, now fast, then slow, breath-taking, almost dangerous. For a few minutes one can live on the edge of danger, but always with the knowledge that safety is only inches and seconds away. Tamer rides are available for the children of all ages who prefer their thrills in more sedate doses. There are beautiful, clean, and true-to-life (better than life?) amusements here also; here everything is pretty, always works, and ends before boredom sets in. There are rides that take one through other countries, fantasy worlds, even into a mildly threatening outer space, and always with the surety of a safe return! Threatening animals become friends, and are totally predictable, clean, and nicer than the real thing. One can even return to the past, seeing of course only nostalgic beauty in the "good old days," and handily passing over any unpleasant memories. The future can be attained in seconds, showing the wonders in store for one as a result of the marvelous technilogical advances of mankind.

Of lesser importance, but still a valid reason for amusement park popularity, is the availability of food of many different types. Cuisine of exotic foreign countries is presented in a fairly reasonable form for a decent price. Where else could be tasted a

bean-paste sweet typical of Japan, a delicate, flaky Napoleon of France, or a foaming cold beer served in a bier haus of Germany? All this, and more, is available at one price, as often as wished.

So are seen two reasons for the tremendous popularity of the amusement parks. All in one package, for one price, instant gratification is there, every day, year-round. All need and desire escape from mundane lives. The amusement parks provide this escape.

ANALYSES OF MODEL ESSAYS

Analysis of 2/1 Model

The essay is not a clear "2" because only the third paragraph is adequately developed; the next-to-the-last sentence of the essay violates the unity and coherence of the paragraph in which it appears; several phrases are unidiomatic; some words and phrases are repeated excessively; the second sentence of the fourth paragraph contains a jarring shift in construction; throughout the essay the point of view vacillates between the first person and the third; and in the first sentence of the last paragraph, "concluding this," a dangling modifier, is particularly confusing because "this" lacks a referent and the phrase is not set off with a comma.

The essay is not a clear "1" because it has a central idea that directly answers the question raised by the topic and that is developed through a clear organizational plan; the transitions are clear, although blatant and conventional; the third paragraph is reasonably coherent, logical, and free from repetition; the essay contains only a few serious grammatical errors, no spelling errors, and no errors in diction that block communication; the syntax is neither consistently rudimentary nor hopelessly tangled; and the essay has an interest-catching opening.

Analysis of 3/2 Model

The essay demonstrates more than the "minimal competence" of a "2" essay, but fails to attain the "definite competence" of a "3." Although the central idea is related to the topic, this idea is not always in clear focus: details, particularly in the second paragraph, describe more the appeals than the effects of advertisements. The opening paragraph has no introduction, merely the thesis divided into two sentences, and the conclusion is a gratuitous restatement of the opening. Transitional phrases are either non-existent or uninspired.

The essay rates above a "2" because it has clear organization, adequate development, and parallel structure. Details are vivid, occasionally novel, and the point of view and tone are generally consistent, the latter being lightly ironic. With the exception of the overuse and misuse of "good" in paragraph three, the extraneous comma in paragraph two, and the necessary comma omitted from paragraph four, word choice is accurate and punctuation correct. Grammatically the essay is altogether sound.

Analysis of 4/3 Model

The essay is not quite a "4" chiefly because the organizational plan is rather ineffective. The second paragraph lacks a clear focus -- given the variety of details contained in it, the writer might very well have gone on to discuss food along with the rides, the animals, and the nostalgic vistas. Of less importance, in the second sentence of the second paragraph, the verb should be nearer its subject; transitional devices are not used skillfully; the writer overuses the "there are" construction in the second paragraph; "technological" is misspelled; and punctuation is sometimes questionable.

The essay is better than a "3" because some of the details are sharply -- or wryly -- observed; the writer turns some nice phrases; the writer manifests a certain sophistication in diction as reflected in the correct use of "sedate," "vicarious," and "gratification"; and the essay contains no grammatical or mechanical errors and only one spelling error.

QUESTIONS AND ANSWERS ON THE RATING OF REGENTS' TEST ESSAYS

(1) Why do we have 2/1, 3/2, and 4/3 models? Why don't we have models of "1," "2," "3," and "4" essays?

All of the discrete ratings cover a wide range of writing performance, particularly the "1." An essay may be assigned a "1" because it is only one sentence long, because it is off the topic, because it contains grammatical errors that frustrate the writer's attempt to communicate, because it is totally lacking in structure, because its points are undeveloped, and so on. There are very, very low "1's," and there are "1's" that are almost passing. While "2," "3," and "4" do not cover so wide a range, it would still be impossible simply to pick one model and say, "This is it." The example would, of necessity, be a low "3," a middling "3," or a high "3." The 4/3, 3/2, and 2/1 models are intended to represent a very fine borderline.

(2) What specifically does the 2/1 model represent?

The essay chosen as a 2/1 model represents the absolute balance point between the "1" and the "2" essay. The committee which selected the essay would hope that, if the 2/1 model essay were rated by fifty raters, it would receive twenty-five "1's" and twenty-five "2's." A tiny nudge could swing the balance either way. It would be a clear "2," if, for example: a few more supporting details were supplied, the diction were more appropriate, the mechanical and grammatical errors were fewer, or the coherence were improved. On the other hand, it would be a clear "1" if it were a trifle weaker in any one of these aspects.

(3) Must an essay have a thesis sentence to pass?

Not necessarily. Although an explicit thesis sentence is perfectly acceptable, and many -- perhaps most -- of our students need one, many a good writer can make the implied thesis clear and can organize the essay well enough so that the reader can follow the line of thought without the writer's having revealed the organizational plan in the introductory paragraph.

(4) Must the essay follow a set formula?

No.

(5) What should be done with essays that are off the topic?

We face two problems here. One involves the student who has a prepared essay and tries to fit it to the topic; the other involves the student who misreads or misunderstands the topic. When raters find an essay that is completely off the topic, they must fail the essay. Misreading is more problematic. Many students who wrote on the topic "Children should never be disciplined by corporal punishment. Defend or attack the statement." thought that corporal punishment was the same as capital punishment. Similarly, a few students who wrote on the topic "Name two or three qualities which you feel a person should possess in order to be a good employee." discussed qualities of a good employer rather than a good employee. When a writer misreads the topic this grossly, the essay should be failed. Most of the misreadings, however, are not so blatant. Many raters found themselves perplexed by the responses to the following two topics: "Discuss the most important moral qualities an elected official should have." and "What qualities of character do you regard as important in a person you would choose as a friend?" Students writing on the latter topic would blithely talk about how their friends should have good looks, an effervescent personality, and

plenty of money more often (or so it seemed) than they would talk about qualities of character such as honesty, integrity, and trustworthiness. Much of the same was true of the former topic, where students would talk about charisma, intelligence, and charm. Seldom, if ever, was an essay totally off the topic: a typical thesis sentence might read "My friends should be loyal, intelligent, honest, and easy to get along with." The raters must penalize the essay for this type of misunderstanding, but such an essay should not be failed out of hand. If the essay is well-written and the student does not seem to be deliberately evading the topic, the essay might well deserve one of the passing scores.

The question of whether the writer can both attack and defend an issue when the topic says "attack or defend" has been raised. When the student deals with pro and con arguments but takes a clear stand on one side of the issue, the answer is definitely yes. Doing so is not merely acceptable, it is meritorious: "although the 55 mph speed limit cost motorists some time and encouraged many citizens to break the law, it should be reinstated because it saved lives, conserved gas, and reduced the number and severity of accidents" is clearly more sophisticated than "the 55 mph speed limit should be reinstated because it saved money, lives, and gasoline." The student who simply attacks and defends without coming down on one side or the other does imperil the chances of passing. However, the student who writes a good fence-straddling essay should be passed.

(6) May the student modify the topic?

Students may make reasonable modifications of the topic. For example, given the topic "What courses that you did not take in high school do you now wish you had taken?" students may state that there are no such courses and explain why. Also, students do not have to discuss specific courses, but may state that they should have taken more courses in an area such as English or history.

Students should not be penalized for narrowing the topic. For example, given a topic which asks for a discussion of the goals of the women's movement, students could narrow the topic by discussing only economic issues.

Students may handle the topic in the first person or the third person, regardless of the person in which the topic is stated. For example, given the topic "Do you agree with the goals of the women's movement?" students may answer, "The goals of the women's movement are valid," and continue in the third person.

(7) How should the rater react to obviously spurious statistics and obviously counterfeit examples?

We must keep in mind that the student writing for the Regents' Test does not have access to an almanac or a set of encyclopedias. Raters should, therefore, be very patient with approximate statistics and with dubious uncles. At the same time, raters must keep in mind that, to the extent examples and statistics are incredible, they are rhetorically ineffective and thus lessen the essay's chances of passing. Writers who say that the accident rate dropped by approximately 10% while the 55 mph speed limit was in effect strengthen their case; writers who say that the accident rate was cut in half while the 55 mph speed limit was in effect weaken theirs.

(8) How should we rate an essay of comic or satiric intent?

Reward the successful and penalize the inept.

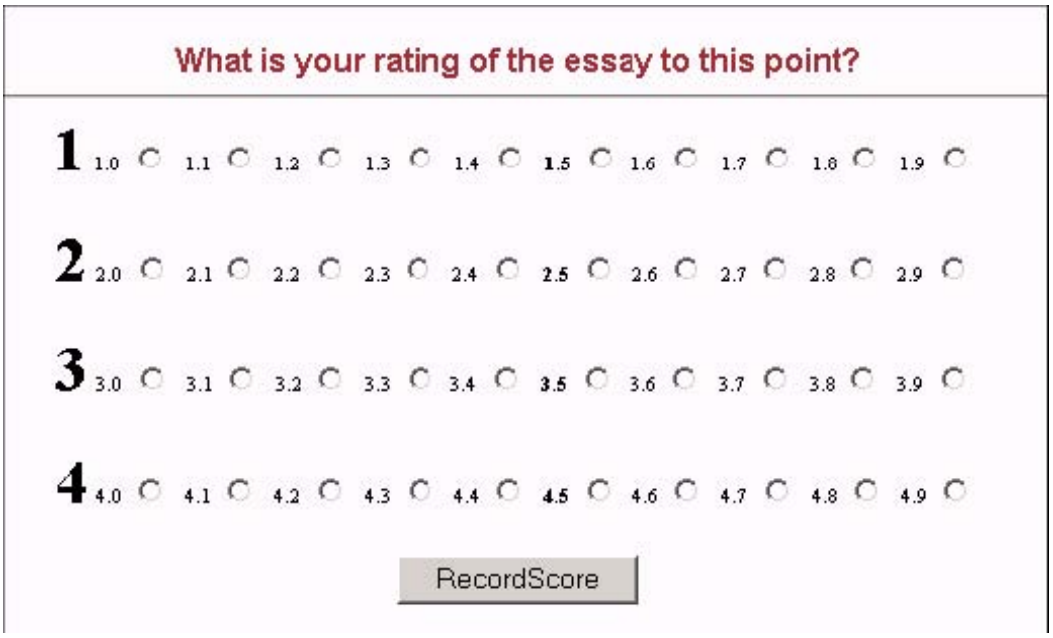
Last updated: September 20, 1999

Appendix B

Prototype for Online Essay Rating Task

Topic # 722: Failure can often teach more than success can. Has failure ever taught you a valuable lesson? Discuss.

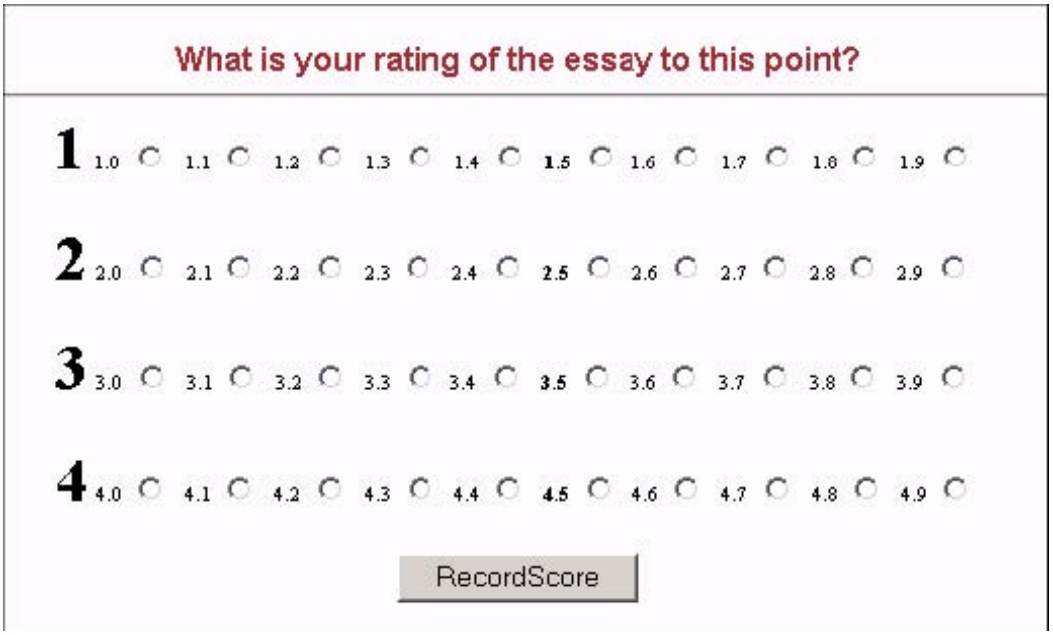
It is true that one can learn from failure in life. Failure in my life has been the result of poor judgement, mistakes in decision-making and lack of goals.



Topic # 722: Failure can often teach more than success can. Has failure ever taught you a valuable lesson? Discuss.

It is true that one can learn from failure in life. Failure in my life has been the result of poor judgement, mistakes in decision-making and lack of goals.

Poor judgement as a teenager caused problems for me because I chose delinquent friends. Often my parents warned me to be careful when I met new people. I did not listen to my parents advice, thinking other people would not influence my actions. My parents were right, after all. I began to stay out late on school nights and steal my parents' car during the night to cruise with these friends. I did not care the next day when I fell asleep at school. Where are these friends now? Some are in jail, some are ill from living a hard, fast teenage life, but one is dead. He died while driving under the influence of alcohol at age twenty. My poor judgement allowed me to believe that I did not need to finish high school. I'd rather be with this group of friends. We all dropped out during our Senior year. Poor judgement cost me a real high school diploma.

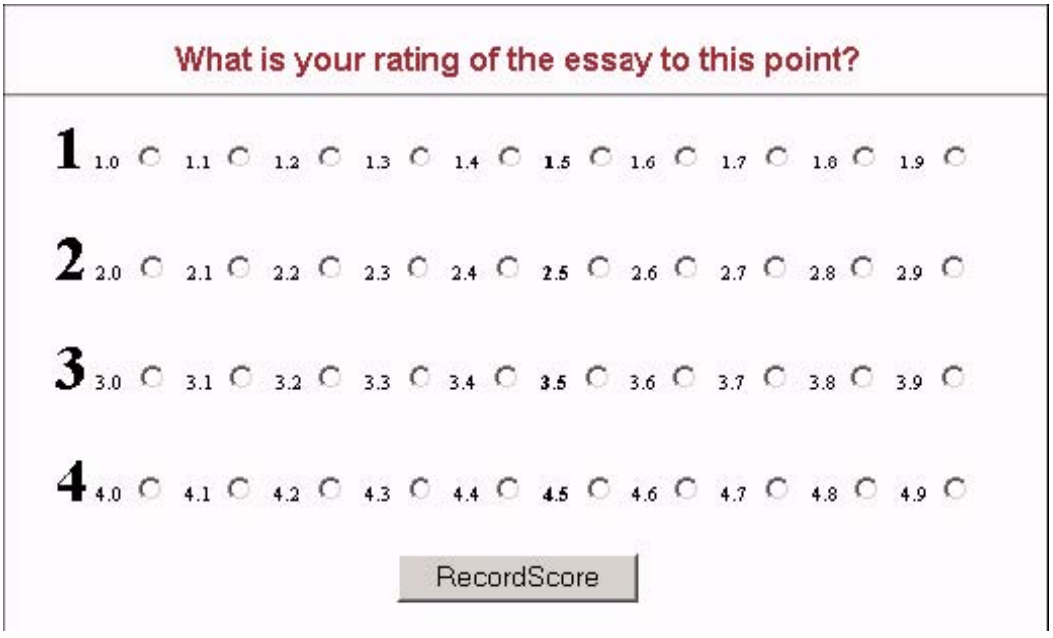


Topic # 722: Failure can often teach more than success can. Has failure ever taught you a valuable lesson? Discuss.

It is true that one can learn from failure in life. Failure in my life has been the result of poor judgement, mistakes in decision-making and lack of goals.

Poor judgement as a teenager caused problems for me because I chose delinquent friends. Often my parents warned me to be careful when I met new people. I did not listen to my parents advice, thinking other people would not influence my actions. My parents were right, after all. I began to stay out late on school nights and steal my parents' car during the night to cruise with these friends. I did not care the next day when I fell asleep at school. Where are these friends now? Some are in jail, some are ill from living a hard, fast teenage life, but one is dead. He died while driving under the influence of alcohol at age twenty. My poor judgement allowed me to believe that I did not need to finish high school. I'd rather be with this group of friends. We all dropped out during our Senior year. Poor judgement cost me a real high school diploma.

Not only did poor judgement lead me to failure, but making mistakes with some of my life's most important decisions, as well. I met a man named Jim who convinced me that he could give me the world on a silver platter. We got married when I was nineteen, against my parents wishes. Jim drank heavily and had a drug addiction. He was old enough to go out to bars without me, so it was also easy for my new husband to continue "playing the field". When I found out and confronted him, he would beat me severely. Two children later, I decided it was time to send this man to jail and divorce him. Few would argue that Jim was a drastic mistake. That silver platter was silver plated.



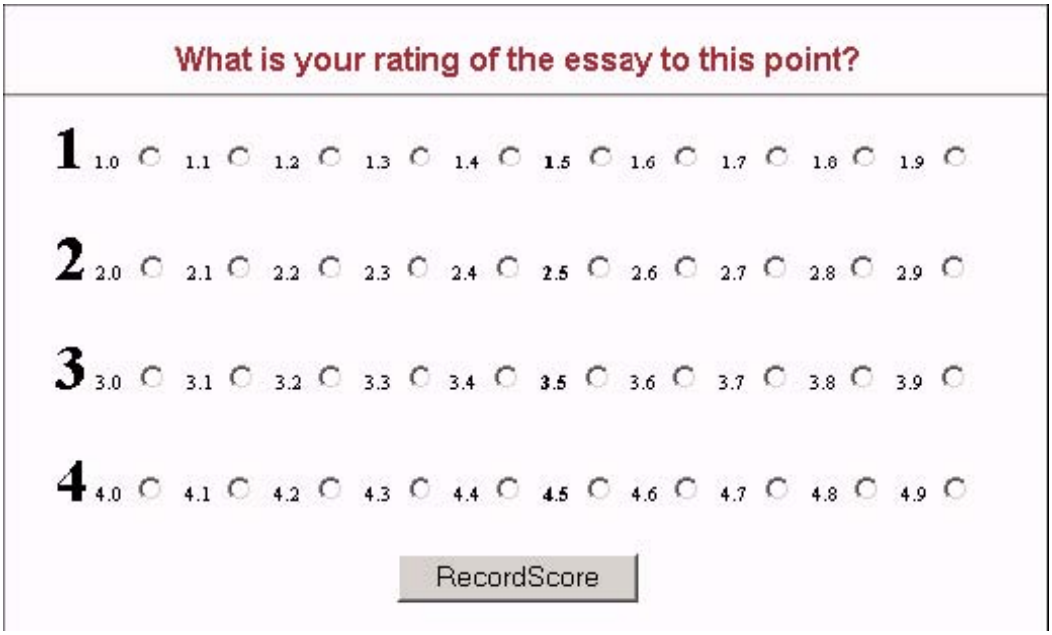
Topic # 722: Failure can often teach more than success can. Has failure ever taught you a valuable lesson? Discuss.

It is true that one can learn from failure in life. Failure in my life has been the result of poor judgement, mistakes in decision-making and lack of goals.

Poor judgement as a teenager caused problems for me because I chose delinquent friends. Often my parents warned me to be careful when I met new people. I did not listen to my parents advice, thinking other people would not influence my actions. My parents were right, after all. I began to stay out late on school nights and steal my parents' car during the night to cruise with these friends. I did not care the next day when I fell asleep at school. Where are these friends now? Some are in jail, some are ill from living a hard, fast teenage life, but one is dead. He died while driving under the influence of alcohol at age twenty. My poor judgement allowed me to believe that I did not need to finish high school. I'd rather be with this group of friends. We all dropped out during our Senior year. Poor judgement cost me a real high school diploma.

Not only did poor judgement lead me to failure, but making mistakes with some of my life's most important decisions, as well. I met a man named Jim who convinced me that he could give me the world on a silver platter. We got married when I was nineteen, against my parents wishes. Jim drank heavily and had a drug addiction. He was old enough to go out to bars without me, so it was also easy for my new husband to continue "playing the field". When I found out and confronted him, he would beat me severely. Two children later, I decided it was time to send this man to jail and divorce him. Few would argue that Jim was a drastic mistake. That silver platter was silver plated.

Failure in my life has come from a lack of goals. I never saw past tomorrow, until I became a single mom with two children to support and very little money or education. I never anticipated, as a teenager, going to college, being realistic in finding someone to spend my life with or how I would obtain the material goods I dreamed I would have some day. Goals make one see beyond tomorrow.



Topic # 722: Failure can often teach more than success can. Has failure ever taught you a valuable lesson? Discuss.

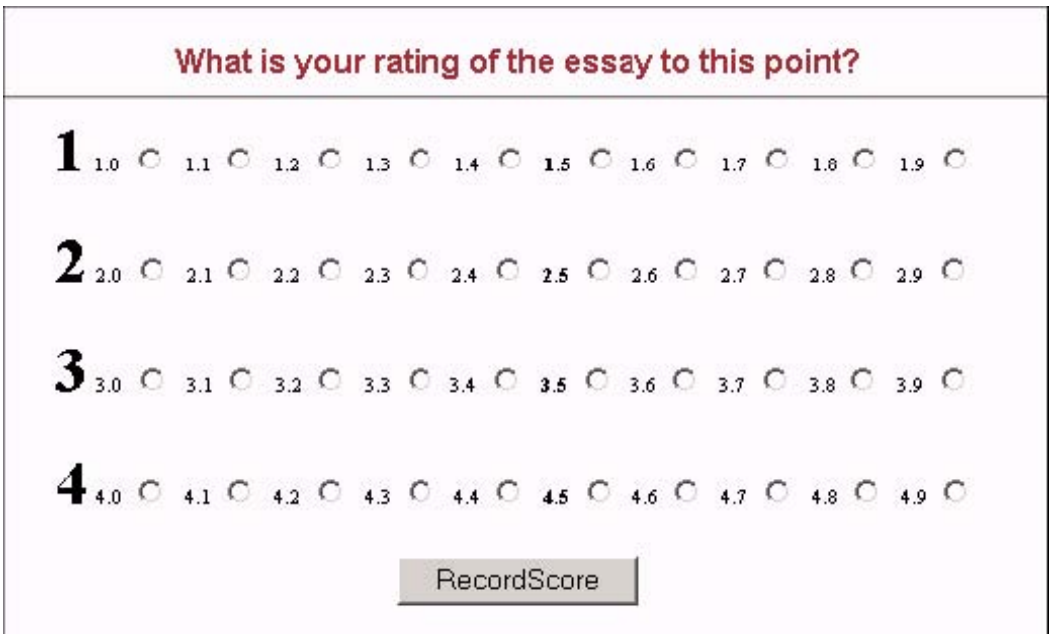
It is true that one can learn from failure in life. Failure in my life has been the result of poor judgement, mistakes in decision-making and lack of goals.

Poor judgement as a teenager caused problems for me because I chose delinquent friends. Often my parents warned me to be careful when I met new people. I did not listen to my parents advice, thinking other people would not influence my actions. My parents were right, after all. I began to stay out late on school nights and steal my parents' car during the night to cruise with these friends. I did not care the next day when I fell asleep at school. Where are these friends now? Some are in jail, some are ill from living a hard, fast teenage life, but one is dead. He died while driving under the influence of alcohol at age twenty. My poor judgement allowed me to believe that I did not need to finish high school. I'd rather be with this group of friends. We all dropped out during our Senior year. Poor judgement cost me a real high school diploma.

Not only did poor judgement lead me to failure, but making mistakes with some of my life's most important decisions, as well. I met a man named Jim who convinced me that he could give me the world on a silver platter. We got married when I was nineteen, against my parents wishes. Jim drank heavily and had a drug addiction. He was old enough to go out to bars without me, so it was also easy for my new husband to continue "playing the field". When I found out and confronted him, he would beat me severely. Two children later, I decided it was time to send this man to jail and divorce him. Few would argue that Jim was a drastic mistake. That silver platter was silver plated.

Failure in my life has come from a lack of goals. I never saw past tomorrow, until I became a single mom with two children to support and very little money or education. I never anticipated, as a teenager, going to college, being realistic in finding someone to spend my life with or how I would obtain the material goods I dreamed I would have some day. Goals make one see beyond tomorrow.

Now, I am thirty years old. I struggle every day to make good judgements calls and cautious decisions. Failure has taught me many lessons, which leads me to keep a constant check on goals I have set for myself. No matter how hard I have to work for the rest of my life, failure has taught me to persevere.


What is your rating of the essay to this point?
1 1.0 1.1 1.2 1.3 1.4 1.5 1.6 1.7 1.8 1.9
2 2.0 2.1 2.2 2.3 2.4 2.5 2.6 2.7 2.8 2.9
3 3.0 3.1 3.2 3.3 3.4 3.5 3.6 3.7 3.8 3.9
4 4.0 4.1 4.2 4.3 4.4 4.5 4.6 4.7 4.8 4.9
RecordScore

Appendix C

Online Study Cell Means from Crossing of 5 Independent Variables Introductory Strategy x Organizational Consistency x Location of Error Density x Replication x Paragraph

Intro	Org	Error	Repl	Para	Mean	Std Dev
1	1	1	1	1	23.0833	3.5792
1	1	1	2	1	25.0833	3.6794
1	1	1	1	2	17.6667	2.7080
1	1	1	2	2	22.4167	3.2039
1	1	1	1	3	18.2500	3.3878
1	1	1	2	3	23.5833	2.9375
1	1	1	1	4	20.2500	3.5452
1	1	1	2	4	24.3333	2.7743
1	1	1	1	5	20.5000	2.9077
1	1	1	2	5	23.9167	2.9064
1	1	2	1	1	26.6667	4.3135
1	1	2	2	1	23.5000	2.9695
1	1	2	1	2	24.6667	4.1851
1	1	2	2	2	23.4167	3.2039
1	1	2	1	3	25.1667	3.9505
1	1	2	2	3	21.8333	3.0401
1	1	2	1	4	25.4167	2.9683
1	1	2	2	4	22.0000	2.5937
1	1	2	1	5	24.6667	3.7254
1	1	2	2	5	22.1667	3.0401
1	2	1	1	1	18.8333	2.5166
1	2	1	2	1	17.9167	3.7040
1	2	1	1	2	17.6667	2.6400
1	2	1	2	2	17.5000	3.0896
1	2	1	1	3	18.1667	3.4597
1	2	1	2	3	18.5833	3.3699
1	2	1	1	4	18.3333	1.9695
1	2	1	2	4	18.7500	3.5452
1	2	1	1	5	18.2500	3.0488
1	2	1	2	5	19.1667	3.3257
1	2	2	1	1	23.1667	3.1286
1	2	2	2	1	23.5833	3.6045
1	2	2	1	2	21.0833	2.7455
1	2	2	2	2	21.0833	3.7769
1	2	2	1	3	23.5000	2.2361
1	2	2	2	3	23.3333	3.3665
1	2	2	1	4	22.1667	3.1575
1	2	2	2	4	23.6667	3.6515
1	2	2	1	5	21.5833	3.2879
1	2	2	2	5	22.3333	3.5760

1	3	1	1	1	22.2500	3.9341
1	3	1	2	1	16.5833	2.5391
1	3	1	1	2	22.0833	4.8703
1	3	1	2	2	19.4167	2.6443
1	3	1	1	3	22.1667	4.7832
1	3	1	2	3	18.8333	3.9734
1	3	1	1	4	22.5833	3.8720
1	3	1	2	4	20.7500	4.1148
1	3	1	1	5	23.1667	4.2391
1	3	1	2	5	19.9167	3.6794
1	3	2	1	1	22.0000	1.9069
1	3	2	2	1	21.0833	3.4499
1	3	2	1	2	23.1667	2.7579
1	3	2	2	2	20.7500	3.5961
1	3	2	1	3	23.6667	2.9949
1	3	2	2	3	19.6667	3.0251
1	3	2	1	4	23.3333	3.7009
1	3	2	2	4	19.8333	3.1286
1	3	2	1	5	22.6667	3.3121
1	3	2	2	5	19.6667	2.7743
2	1	1	1	1	17.5000	4.4210
2	1	1	2	1	25.7500	7.5091
2	1	1	1	2	17.4167	2.4664
2	1	1	2	2	21.4167	5.9461
2	1	1	1	3	18.6667	3.2287
2	1	1	2	3	21.4167	2.6443
2	1	1	1	4	21.0833	4.3788
2	1	1	2	4	23.9167	6.0672
2	1	1	1	5	21.3333	4.2711
2	1	1	2	5	24.0000	6.3102
2	1	2	1	1	23.3333	3.8455
2	1	2	2	1	23.0833	5.0355
2	1	2	1	2	24.3333	3.2004
2	1	2	2	2	22.1667	3.1861
2	1	2	1	3	26.1667	3.8099
2	1	2	2	3	21.5833	2.6785
2	1	2	1	4	21.6667	2.8710
2	1	2	2	4	20.3333	2.6400
2	1	2	1	5	23.3333	3.0551
2	1	2	2	5	20.6667	3.0551
2	2	1	1	1	19.5000	2.6799
2	2	1	2	1	20.5833	2.9987
2	2	1	1	2	20.3333	4.0527
2	2	1	2	2	19.2500	1.9598
2	2	1	1	3	20.5000	3.6307
2	2	1	2	3	21.0000	2.7961

2	2	1	1	4	20.8333	3.5377
2	2	1	2	4	20.5833	2.8749
2	2	1	1	5	20.6667	3.6265
2	2	1	2	5	21.0000	2.7303
2	2	2	1	1	24.5833	3.4499
2	2	2	2	1	23.2500	5.0457
2	2	2	1	2	24.5000	4.3797
2	2	2	2	2	21.9167	4.0104
2	2	2	1	3	23.6667	3.8218
2	2	2	2	3	22.8333	3.7618
2	2	2	1	4	25.4167	4.6015
2	2	2	2	4	21.5000	2.5045
2	2	2	1	5	23.1667	4.0862
2	2	2	2	5	21.6667	3.1431
2	3	1	1	1	24.0833	4.6993
2	3	1	2	1	22.0833	4.9075
2	3	1	1	2	23.0833	3.8485
2	3	1	2	2	19.8333	3.7132
2	3	1	1	3	21.5000	3.8019
2	3	1	2	3	21.9167	4.3580
2	3	1	1	4	22.8333	4.1084
2	3	1	2	4	22.8333	5.1493
2	3	1	1	5	22.2500	3.8406
2	3	1	2	5	22.2500	4.9932
2	3	2	1	1	23.5833	3.8009
2	3	2	2	1	26.7500	3.2787
2	3	2	1	2	25.0000	2.7634
2	3	2	2	2	27.1667	3.7618
2	3	2	1	3	25.8333	3.8099
2	3	2	2	3	27.3333	3.7009
2	3	2	1	4	25.1667	3.8099
2	3	2	2	4	25.7500	3.4411
2	3	2	1	5	24.8333	3.5119
2	3	2	2	5	26.2500	3.6213

Appendix D

Summary Table, ANOVA for Study 2
Introductory Strategy x Organizational Consistency x Location of Error Density x Replication x Paragraph

	SS	Df	MS	F	p-value
Intro	437.80	1	437.80	7.73	0.018
Org	629.72	2	314.86	9.51	0.001
Error	2235.02	1	2235.02	35.54	0.000
Para	121.20	4	30.30	3.51	0.014
Rep	46.94	1	46.94	1.54	0.254
Intro x Org	903.78	2	451.89	12.84	0.000
Intro x Err	6.14	1	6.14	0.22	0.651
Intro x Para	10.77	4	2.69	0.94	0.447
Intro x Rep	45.51	1	45.51	1.82	0.205
Org x Error	208.01	2	104.01	2.13	0.143
Org x Para	141.11	8	17.64	2.97	0.005
Org x Rep	318.84	2	159.42	5.32	0.013
Error x Para	243.38	4	60.85	10.46	0.000
Error x Rep	368.04	1	368.04	19.06	0.001
Para x Rep	15.74	4	3.94	1.27	0.294
Intro x Org x Error	303.09	2	151.55	3.35	0.054
Intro x Org x Para	105.27	8	13.16	3.47	0.002
Intro x Org x Rep	286.18	2	143.09	3.82	0.038
Intro x Error x Para	27.40	4	6.85	2.37	0.067
Intro x Error x Rep	1.34	1	1.34	0.03	0.863
Intro x Para x Rep	112.55	4	28.14	9.36	0.000
Org x Error x Para	152.41	8	19.05	4.89	0.000
Org x Error x Rep	964.28	2	482.14	11.25	0.000
Org x Para x Rep	53.93	8	6.74	2.02	0.054
Error x Para x Rep	47.16	4	11.79	3.29	0.019
Intro x Org x Error x Para	114.33	8	14.29	3.48	0.002
Intro x Org x Error x Rep	62.48	2	31.24	1.37	0.276
Intro x Org x Para x Rep	103.42	8	12.93	3.75	0.001
Intro x Error x Para x Rep	34.95	4	8.74	3.27	0.020
Org x Error x Para x Rep	87.81	8	10.98	3.23	0.003
Intro x Org x Error x Para x Rep	38.38	8	4.80	1.14	0.346

Appendix E

INSTRUCTIONS FOR SCORING PORTFOLIOS

DESCRIPTION OF PORTFOLIO SCORING PROCEDURE

Raters should read each portfolio quickly to gain a general impression of its overall quality. Essays within a portfolio should be read in the order in which they are presented from first to last. However, raters should read quickly to evaluate the entire portfolio holistically and assign one rating to the portfolio as a unit. For this assessment, the raters' task is to assess the quality of the student's writing based on the portfolio as a whole.

Raters should suspend judgment about the quality of a portfolio until they have finished reading the entire portfolio.

The portfolios are to be rated on the following four-point scale in which "1" is the lowest score and "4" is the highest score. When judging the quality of any given portfolio, raters should assign the rating point associated with the description below that best fits their assessment of that portfolio.

4: The "4" portfolio contains writing with clear central ideas that relate directly to the assigned topics. The writing is characterized by clear organizational plans. Major points are developed logically and are supported with concrete, specific evidence or details that arouse the reader's interest. In general, the writing reveals the student's ability to select effective, appropriate words and phrases; to write varied, sophisticated sentences; to make careful use of effective transitional devices; and to maintain a consistent, appropriate tone. The writing is essentially free from mechanical errors, it contains no serious grammatical errors, and ideas are expressed freshly and vividly.

3: The "3" portfolio contains writing with clear central ideas relating directly to the assigned topics. The portfolio generally differs from a "4" portfolio in that it shows definite competence, but lacks distinction. The writer's examples and details are pertinent, but may not be particularly vivid or sharply observed; the word choice is generally accurate, but seldom--if ever--really felicitous. The writer characteristically adopts an appropriate, consistent tone. The writing may contain a few errors in grammar and mechanics.

2: The "2" portfolio meets only the basic criteria, and those in a minimal way. Central ideas are directly related to the assigned topics and presented with sufficient clarity that the reader is aware of the writer's purpose. The writing generally reflects the student's ability to organize writing clearly enough for the reader to perceive his or her plan and to compose text that coherently presents some evidence or details to substantiate the points. The writer uses ordinary, everyday words accurately and idiomatically and generally avoids both the monotony

created by series of choppy, simple sentences and the incoherence caused by long, tangled sentences. Although the writing may contain a few serious grammatical errors and several mechanical errors, they are not of sufficient severity or frequency to obscure the sense of what the writer is saying.

1: The "1" portfolio characteristically reflects the writer's lack of competence in controlling any one of the following problems to an extraordinary degree or several of them to a limited degree: the lack of central idea; the lack of clear organizational plan; points which are not developed or are developed in a repetitious, incoherent, or illogical way; the lack of direct relation between text and assigned topic; an excessive number of serious grammatical errors or mechanical errors; inaccurate or unidiomatic use of ordinary, everyday words; a limited vocabulary such that the words chosen frequently do not serve the writer's purpose; frequently rudimentary or tangled syntax. Or the portfolio is so brief that the rater cannot make an accurate judgment of the writer's ability.

Appendix F

Training Portfolios

Portfolio #219 Homogeneous 1" Essays

Portfolio #467 Homogeneous 2" Essays

Portfolio #350 Homogeneous 3" Essays

Portfolio #442 Homogeneous 4" Essays

PASTIME.33
SS# _____

EXTEMPORANEOUS

Portfolio #219 Essay # _____

Topic: "Name your favorite pastime and explain why you enjoy it."

My favorite past time is camping. One of the reasons why would have to be because I love spending time with my family. Another reason pertains to the enjoyment I recieve from fishing in the lake by our campgrounds. Last but not least, my whole family enjoys swimming.

The most important reason for me to go camping is to spend time with my family. We play games such as volleyball or sit around playing cards. One of my favorite games is Gin Rummy. At night while the camp fire is blazing we catch up on gossip or whats been happening.

The second activity while camping is fishing. The peacefulness of Lake Laneir will let your mind drift. One time when I went ocean fishing, I caught a blow fish and a hammer head shark. Catching the shark scarred me half to death. The best reason to fish is you get to eat what you catch.

The last reason I like camping is the pleasure of going swimming. My main reason for going swimming is to cool off. My son likes me to throw him in the water, which wears me out! It's also relaxing and after a day of swimming you will have a good nights sleep.

Camping has many great benefits. It is a real experience especially if you have never been. After a weekend of camping you come home relaxed and refreshed.

SOME1ELS.21
SS# _____

Portfolio #219 Essay # _____

Topic: "If you could wake up tomorrow as someone else, who would it be and why?"

If I could chose who to be, it would have to be Dolly Parton. I have admired her since I was a child. One reason is because she's a great country singer. Another reason pertains to her great body. The last reason deals with her acting career.

Dolly Parton is one of the greatest country singers of all time. One of my favorite songs of hers is called I Will Always Love You. Another favorite is Islands in the Stream, which she sings with Kenny Rogers. Not only does she sing with Kenny, but she also sings with Billy Rae Cyrus and others. Dolly Parton is the type of singer who can move her audience. Not all singers can do that, which is why I wish I could sing like her.

Another reason I wish I could be Dolly Parton pertains to her great body. At first, she had big breast, so she went to a doctor to make them smaller. By accident, the doctor made them larger. Dolly sued the doctor, but was left with huge breast. Not only does she have big breast, she's slender, which makes her wanted by every man in the world.

The last reason deals with her acting career. Dolly Parton is one of the most terrific actresses. She has acted in many movies and television shows. There are three movies that are my favorite. One is called the Biggest Little Whore House in Texas. Another movie is Smokey Mountain Christmas. The last movie is Nine to Five. Dolly can play any character in any movie and make it terrific.

Dolly Parton has everything a woman could dream of. She has a great voice, looks, and a great acting career. Now if you had the chance to be her, wouldn't you?

STATSYMB.03
SS# _____

Portfolio #219 Essay # _____

Topic: "Discuss some of the status symbols of today's society."

In today's society there are lots of status symbols. One example is the size of your house and the neighborhood you live in. Another pertains to the make of your car. And a third deals with the clothes you wear such as the brand or the stores where you bought them.

The upper class lives on the hill in beautiful, fancy houses with everything perfect. The carpet, furniture, even the wall paintings have to be the very best from Macy's and BJ White. The cars they drive have the softest leather money can buy. The clothes they buy have to be made by the most famous designers such as Liz Clayborn and Ralph Lauren.

The middle class are happy living in comfortable houses. The carpet and furniture bought little by little of course. The cars they drive, are good on gas. Not to expensive and not a piece of junk. They shop for clothes at WalMart or maybe take a ride to Target.

The lower class are happy with a roof over their head that doesn't leak. If they are lucky they will have carpet that doesn't have holes or isn't worn out. Furniture found on the side of the road that nobody wanted anymore would be like getting something new. They pray every time they get into their old beat up car that it will start. They buy their clothes at Goodwill.

In conclusion, I think today's society puts to much priority on having material things. Each person should experience how the other person lives. They would appreciate more how hard every person has to work for the status symbols they have.

PASTIME.31
SS# _____

EXTEMPORANEOUS

Portfolio #467 Essay # _____

Topic: "Name your favorite pastime and explain why you enjoy it."

My favorite pastime is driving around in my car. The three main reasons driving is my favorite pastime are (1) it helps me relieve my stress, (2) it helps me communicate with my friends, especially the one I care the most about, Kelly Adams, and (3) it also helps me think and get my life back into shape.

The way driving helps me to relieve stress is that it puts me in a relax state of mind. It also helps me forget about all the bad things in life. For example, when I am stressed out from school and work, I just jump into my Z71 Cheverolete pick-up truck and drive until I feel calm. Feeling the purr of the big engine has a way of reassuring me that everything is going to be alright. A long drive in my truck gives me a feeling that my life isn't going to the dogs completely.

Driving helps a whole lot when it comes down to communicaton for me. Without driving, I couldn't see my girlfriend Kelly Adams or any of my friends. Driving is especially important on the weekends because that is when everybody cruises. All night we ride up and down Washington Rd. and talk to everyone we know or see. Some people say we cause trouble. But we don't. This is the only thing people can do for entertainment in Augusta if they are under 21.

Finally, driving helps me think about school, work and my life as a whole. When I drive to think, I drive down old country roads such as the one in Hephzibah. They hardly ever have cars going down the streets there and the trees grow over the pavement. It is real dark even when the sun is at its brightest. I like to drive down it and park on the side of the street, then I imagine the old cars going down it and wonder where my life is really headed.

Driving helps me in many ways. It helps me get my life straight by giving me time to think without interruptions. It helps me communicate with my friends. And, it especially helps me to be near the one I care about the most, Kelly Adams.

SOME1ELS.03
SS# _____

Portfolio #467 Essay # _____

Topic: "If you could wake up tomorrow as someone else, who would it be and why?"

If I could wake up tomorrow as someone else, it would be my father, Paul Michael Hamilton. The reason why I would like to be him is because he is the greatest person in my life, and he is a fabulous role model. He shows responsibility, love and is highly respected. When life gets tough, he seems to be able to solve life's problems. He always has a solution for every crisis that arises.

The one trait that is most evident in him is definitely responsibility. He is always on time, and he accomplishes everything that he sets out to do. When he sets a date for a goal to be achieved, he will make sure that his job is done. He always takes full responsibility for everything that he does. He tries to teach me responsibility as I begin my life as an adult. When I backed out of a commitment recently to help a friend lay the foundation for his house, my dad reminded me that a man's word is his pledge. Needless to say, I followed through on my promise to my friend. Responsibility is the most meaningful characteristic about my father.

The second trait that is visible in my father is love. He shows it in many distinct ways. He reveals love by stepping into my life and cheering me up, to sitting down and having serious talks with me. As father and son, love is prevalent when we spend time together. He strives to lead me in the right direction. He shows me that without love, where would our family be?

Accomplishment is the third trait that stands out in my father. I respect him in many ways. Beginning with the three companies that he owns and runs all at one time, to the quality time he spends with his family at home and at the lake traveling on the two yachts that he owns. I respect him for the wonderful family that he and my mother have built and for the good times he brings into our home. He has made me the person that I am today, and that is what I respect about him the most.

These three traits make me think about waking up tomorrow and being my father. Today, I have already started to be like him by working with his companies, and following the advice that he gives me. Someday I hope to be just like him, because he is an inspiration to me, and to everyone that he encounters. I feel that he is very successful, and I want to walk in his footsteps one day.

STATSYMB.09
SS# _____

Portfolio #467 Essay # _____

Topic: "Discuss some of the status symbols of today's society."

Today's society is surrounded by status symbols. America is one of the biggest money hungry countries in the world. Commercials on television make it look like you have to have certain things just to meet the standards of our society. Several status symbols include what you wear, what kind of car you drive, and where you live.

The biggest status symbol is what you wear. If you see someone wearing a rolex watch, polo pants, shirt, and shoes you would draw the conclusion he or she has money. However, if you see someone with a pair of ripped jeans, a dirty t-shirt, and a old pair of tennis shoes you would think that they are very poor. Most people don't realize that no matter what someone wears, you can't tell how much money they have.

Along with clothing, the type of car you drive also plays a big role in how you are judged by others. Everyone has a dream car. It might be an old mustang or a new Mercedes 300E. Advertisements want us to believe that a car can fullfill someone's self-worth. People have a tendency to compete with each other. If your neighbor buys a new car and you have a clunker, you may be influenced to buy a new car that's even better, just so you will look good.

Where you live also is a status symbol. Living in Beverly Hills and living in lower New York are on opposite ends of the spectrum. People would look up to those who live in Beverly Hills because we see them as having power and money. However, people would not show respect to those who live in lower New York because they don't have the same material possessions.

These are just several status symbols that people are defined by. People need to learn to look past what someone looks like or where they live before we judge them and understand that power and money does not justify someone's inner self.

PASTIME.JR
SS# _____

EXTEMPORANEOUS

Portfolio #350 Essay # _____

Topic: "Name your favorite pastime and explain why you enjoy it."

Gardening. Rowing. Reading. Cross-stitching. Long country drives. These are all activities I enjoy. However, I can not name one of them as my favorite. Instead, I believe I have more unconventional favorite pastime. I love to spend my time with my family! In today's world, this does not appear to be the most popular way to spend one's freetime, however I find it to be most advantageous. Spending free time with my family provides me with opportunity to grow closer to them; it enables me to still participate in other activities I enjoy and finally it gives me the freedom to be myself.

I have found the more fun time I spend with my family, the closer we become. The daily grind can become a frustrating way of life and I find it too easy to take it out on those I love the most. Instead, we have come to take advantage of the spare time we have together. We find we work together well as a team and grow closer in the process. We do anything from lay around in the house watching cartoons in our jammies to volunteer projects that help us all, like building a deck and a raised garden in the backyard. This brings me to my next point.

Not only do I find myself closer to my family in my spare tie, but I find I still have the opportunities to do other things I love. While my husband and children watch a movie they chose together at Movie Gallery, I'll curl up on the sofa and catch up on some cross-stitch. While my little boy plays catch with my husband, my daughter and I plant vincas and coleus in the garden, my pride and joy. We may not all be working on the same project, but sometimes just being in the same room or in the backyard together is all that is needed.

The most important aspect of spending time with my family for me is that I can just be myself. This is the one group of people I feel most comfortable with. They love me for me. We can try any new adventure out together and still come out with giggles and hugs and kisses. Most of all, I can try anything. Whether I succeed or not, they are always there to cheer me on.

My favorite pastime is my family. I have more fun with my husband, son, and daughter than anyone else in the world. My favorite and most cherished memories involve these three people. By spending those few extra moments together, we have become closer, learned new things trying out each other's adventures, and gained the opportunity to unconditionally be ourselves. I think my pastime is rare in this day and age, but I consider myself lucky to have such a special hobby.

SOME1ELS.JR
SS# _____

Portfolio #350 Essay # _____

Topic: "If you could wake up tomorrow as someone else, who would it be and why?"

I do not want to be famous. I do not want to become someone else and be known for his or her great ideas. Instead, I want to be known for my own accomplishments. I do however, wish that I had the guts to be like some of the great leaders of our time. If I could wake up tomorrow morning as someone else, I would wake up as myself, only with a different view on life. I watch other people do what I dream of doing. Tomorrow morning, I want to become the kind of person who does instead of observes.

The first step on my way to becoming the new me is to wake up with poise. I want to walk with the air of self confidence that I see in James Earl Jones, Madeline Albright or Jackie Kennedy Onassis. I want to be graceful and carry myself with my head held high and a sense of assurance.

Tomorrow morning I want to wake up without any inhibitions. I want to take more risks and not be so scared to try new things. I want to stand up and fight for what I believe in and not hold back because I am afraid of what others might think of me. I want to fight for what is right in this world with Steven Biko's bravery and Martin Luther King's insight. Tomorrow I want to become a new me. A me who speaks her voice when it is needed instead of sitting in the corner and not saying anything at all. I do not want to allow things to happen, rather I want to make them happen.

Finally, I must wake up with joy. I need to stop focusing on negative aspects of life because dwelling on those distractions wastes precious time and energy. I need to quit hating and start loving. I need to realize that time is short and there are too many good things in my life to waste another moment dwelling on negative issues. I need to get outside and smell the sunshine to appreciate the potential I have to make a difference.

I do not want to live someone else's adventures. We can only know so much about others. To be them, we must be all of them, both the good and the bad. I want to be all of ME. I want to stand out for who I am and for what I have accomplished. If I could wake up tomorrow morning as someone else, I would become the me who does instead of watches.

STATSYMB.JR
SS#_____

Portfolio #350 Essay#_____

Topic: "Discuss some of the status symbols of today's society."

Tommy Hilfiger, Mercedes Benz, Hilton Head and the game of golf all have one thing in common. Each may be regarded as a status symbol. Our society is filled with people longing to be thought of as bigger, better, more powerful and of course richer than anyone else. By wearing name brand clothing, driving fancy cars, vacationing at luxurious and expensive resorts and playing a game originally reserved for the elite, many people feel that they are attaining recognition.

I fell into the status symbol trap without even knowing it. I tried to make a conscientious effort to avoid becoming what others perceived as cool. Before I knew it, however, I was buying all my clothes from Gap and Banana Republic. I wouldn't be caught dead in a Wal-Mart store, and my shoes had to cost at least fifty dollars. I also found myself bragging about the price of the clothes and shoes that I wore. I wanted other people to think that these things symbolized my worth, even though I really could not afford most of the things I bought.

The biggest (and most expensive!) status symbol I fell for was a car. For some reason, I thought Volkswagons were what cool rich people drove. Although we were both starving college students, my husband and I bought the car that we thought would make us look really cool- a brand new 1996 Volkswagon Jetta Trek. A limited Edition. The last one in the state, according to the salesman. My ego went up about ten notches that day as I drove off in my shiny new status symbol.

For the last two years, I have been dressing well and driving a really neat car. However, my eagerness to show off all of my cool stuff started to bite me back. To this day, I am still paying for clothes I do not even wear anymore. And that awesome car? I still drive it, but newer, neater cars have since come out and I have three more years to go on my lease. I am also stuck with payments that prevent me from buying anymore cool clothes.

I have recently discovered the art of garage-saling and have found that you can still shop at the Gap through someone else's old clothes and spend a lot less money. Nobody knows that I only spent fifty cents for the shirt but the label still says Gap. The moral of the story is that status symbols do not necessarily maintain their value.

Today I see people walk around in fancy clothes and drive off in luxurious cars. I know most of them really do not have the money that their possessions are trying to portray. Status symbols are merely objects that we use to try and impress other people. Unfortunately, those symbols can get us in a lot of trouble if we let our greed to be cool in the eyes of others become our goal.

PASTIME.19
SS# _____

EXTEMPORANEOUS

Portfolio #442 Essay # _____

Topic: "Name your favorite pastime and explain why you enjoy it."

My favorite activity is watching movies and ferreting out their meaning. I enjoy all types, and sometimes frustrate my family and friends, who all say that I will watch anything. They accuse me of having no taste just because I can watch three Rambo movies in a single day. I, on the other hand, insist that Stallone is just too subtle for most people. Pearls before swine, as it says in the Bible.

From my perspective, the Rambo saga portrays an American ideal. Consider the scene in Rambo II in which his Vietnamese girlfriend, who barely speaks English, speaks at some length about his return to America and her desire to go with him. Rambo's response is classically understated: Yeah, yeah. Just when happiness seems within Rambo's reach, the dirty Viet Cong round a bend in the river and squeeze a few rounds into his woman, evoking a heart-rending No, no from Sly the master. Shakespeare might have had more to say, but he couldn't have said it any better. Clearly Rambo is from Palooka-ville, U.S.A., and just as clearly, Palooka-ville is one densely populated city.

A more meditative depiction of the American character is that of Stephen Sagal. Like Caine on Kung Fu, he practices the non-violent philosophy of Buddhism, and, also like Caine, he is forced to kick butt just about everywhere he goes. Segal has added a new twist to kicking butt in his last two movies. In both On Deadly Ground and Fire Down Below, Segal plays a tough guy with a soft heart for his Mother Earth. He is a killer with a message, in other words. Whether you're drilling for oil or mining for coal, you must do it responsibly or face the consequences.

There is one cinematic puzzle I can't resolve, though--the James Bond paradox. Why would Goldfinger strap James Bond to a table with a laser meant to saw him in half when a twenty five cent bullet applied immediately to the back of the head would save him untold heartache and misery? Why does Largo leave it to a swimming pool of sharks to dispose of the indestructable Mr. Bond when he has a stable full of thugs at his disposal? Sure, thugs are dumb and messy, but compared to sharks they're brain surgeons. If I had all that time and money invested in a super-criminal hideout and a super-criminal plan to bring the world to its knees, James Bond wouldn't live ten seconds once I got hold of him.

I guess maybe my friends are right. I will watch anything. Maybe, for me, a flickering screen is like someone else's traffic accident. I can't not look. Perhaps it's just too flashy to ignore, like the circus or a drag race, or maybe I'm just a little on the Crazy Side. The truth is that it's probably all of these.

SOME1ELS.30
SS# _____

Portfolio #442 Essay # _____

Topic: "If you could wake up tomorrow as someone else, who would it be and why?"

The prospect of suddenly becoming anyone that I would like to be is one that sets my mind reeling. There are so many considerations that come into play. The responsible thing would be to select someone like Mother Teresa because the world has the greatest need for such people, but who wouldn't rather be Michael Jordan or Eddie Vedder? Indeed, there are times it might be enough just to be one of the Spice Girls. Honestly though, I would choose to be myself -- with a bottomless bank account.

I think that money is the only thing that gives one the freedom to make finding out who you are a full time endeavor. Am I a traveling man, or am I a blues man? Do I have personality plus, or am I merely spirited? Am I moving along the road to Nirvana or not? Just how charitable can I be? If you have a lot of money and you give a lot of money away, but you're still left with a lot of money, what have you really proven? These are questions I could really enjoy answering.

I have to admit that I'd face serious challenges to my self-concept. With the same fervor that I pray for God to smite down my enemies, I pray to wake up stinking rich the next morning. I feel that my character sorely needs this kind of testing. If life is a test, and I hear frequently that it is, then let me say now that I have studied enough. I am ready for the final exam. I'm willing to risk damnation to prove myself. So why isn't the cash on my doorstep?

I would share my wealth, really I would. My friends, who already think I'm a hell of a guy, would come to think of me as *one hell of a guy*. Jeep Cherokees and surround sound for everybody. My family, who have been tut-tutting and scratching their heads about me for years, would suddenly say, Why, Richard has really come along lately. I guess he was just a late bloomer. Why deny my family the chance to fine-tune their love for me? Why, Lord?

Yes, I truly think it would be best for all concerned if I were suddenly a rich man. Consider me a volunteer in a raging spiritual war. I know that the temptation will be fierce, but I am eager to face the challenge. You, the reader must think me Christ-like, but consider this. I know you'd do the same for me.

STATSYMB.31
SS# _____

Portfolio #442 Essay # _____

Topic: "Discuss some of the status symbols of today's society."

Status symbols in America come in every shape and size. In some corners of the United States, a bluetick hound is a much prized status symbol. There have been times in history that a chicken in the pot marked you as a prosperous person. For some people today, though, whole corporations are no more than status symbols. Among college students, there is a definite set of prescribed possessions.

The first and most critical status symbol for a college student is the type of car one has. There are certain acceptable variations, and any of them might work. But to attain high status, you must have one of three types of vehicles.

Probably the most commonly prized vehicle among my peers is the great big pickup truck. To be a real success, you should invest in as much chrome as the law will allow and a stereo that will ring doorbells thirty miles away. A truck box adds status as well, but only if you don't keep any tools in it. It is a good idea to have four wheel drive, and you should go mudding from time to time, but you should only let your truck stay dirty long enough for a few people to see it. You must wash it in a timely manner. Better yet, have it washed in a timely manner.

Another acceptable type of vehicle is the cute car. A good example of this would be a pink convertible of any make or model. A pink convertible Karmann Ghia. A 1962 pink convertible Ford Fairlane. The cute car category allows one to select any number of foreign models, the more offbeat the better. Imagine yourself in a Citroen or a Studebaker. The advantage of this strategy is that it allows for more freedom of expression. The disadvantage is that these vehicles tend to leave their drivers walking home or catching a ride in someone else's status symbol, and it requires a special the hell with everything kind of attitude to pull that off. If you can't envision yourself enduring this indignity, better go with option one or option three.

The other high-status vehicle for college students is the large ticket car. Any car that a stockbroker would be proud to be seen in will also work for you. BMW's do nicely, as do Lexus's. At the lowest end of the spectrum is the Honda Prelude or Mazda Miata, which incidentally, bridges the gap between pure status vehicle and cute car. You will get points for owning any of these status symbols, but be careful. Moms and Dads keep a close watch on you when you drive one of these, and are quick to revoke your privileges at the slightest provocation.

I hope that this has been of some help to the reader who wishes to distinguish himself or herself with a car. Let this essay be your guide, and you can't go wrong. If you step outside of these guidelines, you must do so with flair, which requires imagination. Which makes it a very low probability strategy in any age group. Finally--and I can't stress this enough--stay away from Consumer Reports.

Appendix G

RATER SCORING FORM FOR TRAINING PORTFOLIOS

Your Rater Number: _____

FOR EACH PORTFOLIO,

(1) RECORD THE 3-DIGIT PORTFOLIO NUMBER &

(2) INDICATE YOUR RATING OF EACH PORTFOLIO BY CIRCLING THE APPROPRIATE NUMBER.

IMPORTANT: PLEASE DO NOT DISCUSS THE PORTFOLIOS OR YOUR RATINGS WITH ANYONE ELSE. THANK YOU.

PORTFOLIO # __ __ __ RATING 1 2 3 4

PORTFOLIO # __ __ __ RATING 1 2 3 4

PORTFOLIO # __ __ __ RATING 1 2 3 4

PORTFOLIO # __ __ __ RATING 1 2 3 4

Appendix H

RATER SCORING FORM FOR PORTFOLIOS

Your Rater Number: _____

FOR EACH PORTFOLIO,

(1) RECORD THE 3-DIGIT PORTFOLIO NUMBER &

(2) INDICATE YOUR RATING OF EACH PORTFOLIO BY CIRCLING THE APPROPRIATE NUMBER.

IMPORTANT: PLEASE DO NOT DISCUSS THE PORTFOLIOS OR YOUR RATINGS WITH ANYONE ELSE. THANK YOU.

PORTFOLIO # __ __ __ RATING 1 2 3 4

PORTFOLIO # __ __ __ RATING 1 2 3 4

PORTFOLIO # __ __ __ RATING 1 2 3 4

PORTFOLIO # __ __ __ RATING 1 2 3 4

PORTFOLIO # __ __ __ RATING 1 2 3 4

PORTFOLIO # __ __ __ RATING 1 2 3 4

PORTFOLIO # __ __ __ RATING 1 2 3 4

PORTFOLIO # __ __ __ RATING 1 2 3 4

PORTFOLIO # __ __ __ RATING 1 2 3 4

PORTFOLIO # __ __ __ RATING 1 2 3 4

PORTFOLIO # __ __ __ RATING 1 2 3 4

PORTFOLIO # __ __ __ RATING 1 2 3 4

PORTFOLIO # __ __ __ RATING 1 2 3 4

PORTFOLIO # __ __ __ RATING 1 2 3 4

Appendix I

Portfolio Study Cell Means
Order x Portfolio Type x Extemporaneous Essay Position

Order	Port Type	Extemp Pos	Mean	Std. Deviation	N
1	Homog 1	1	1.00	.000	3
		2	1.00	.000	3
		3	1.33	.577	3
		Total	1.11	.333	9
	Homog 2	1	2.00	.000	3
		2	2.33	.577	3
		3	2.00	.000	3
		Total	2.11	.333	9
	Homog 3	1	2.67	.577	3
		2	2.00	.000	3
		3	3.00	.000	3
		Total	2.56	.527	9
	Low Diss	1	2.00	.000	3
		2	2.67	.577	3
		3	2.33	.577	3
		Total	2.33	.500	9
	High Diss	1	2.00	.000	3
		2	2.00	.000	3
		3	1.00	.000	3
		Total	1.67	.500	9
	Total	1	1.93	.594	15
		2	2.00	.655	15
		3	1.93	.799	15
		Total	1.96	.673	45
2	Homog1	1	1.33	.577	3
		2	1.33	.577	3
		3	1.00	.000	3
		Total	1.22	.441	9
	Homog2	1	2.00	.000	3
		2	2.00	.000	3
		3	2.33	.577	3
		Total	2.11	.333	9
	Homog3	1	3.67	.577	3
		2	3.33	.577	3
		3	3.00	1.000	3

		Total	3.33	.707	9
	Low Diss	1	2.33	.577	3
		2	2.33	.577	3
		3	2.00	.000	3
		Total	2.22	.441	9
	High Diss	1	2.33	.577	3
		2	2.00	.000	3
		3	2.00	1.000	3
		Total	2.11	.601	9
	Total	1	2.33	.900	15
		2	2.20	.775	15
		3	2.07	.884	15
		Total	2.20	.842	45
3	Homog1	1	1.33	.577	3
		2	1.00	.000	3
		3	1.00	.000	3
		Total	1.11	.333	9
	Homog2	1	2.00	.000	3
		2	2.00	.000	3
		3	2.33	.577	3
		Total	2.11	.333	9
	Homog3	1	2.33	.577	3
		2	3.33	.577	3
		3	2.33	.577	3
		Total	2.67	.707	9
	Low Diss	1	2.00	.000	3
		2	2.00	1.000	3
		3	2.00	.000	3
		Total	2.00	.500	9
	High Diss	1	1.33	.577	3
		2	1.67	.577	3
		3	2.00	1.000	3
		Total	1.67	.707	9
	Total	1	1.80	.561	15
		2	2.00	.926	15
		3	1.93	.704	15
		Total	1.91	.733	45
4	Homog1	1	1.00	.000	3
		2	1.67	.577	3
		3	1.33	.577	3
		Total	1.33	.500	9

	Homog2	1	2.00	.000	3
		2	2.00	.000	3
		3	2.33	.577	3
		Total	2.11	.333	9
	Homog3	1	2.67	.577	3
		2	2.67	.577	3
		3	2.67	.577	3
		Total	2.67	.500	9
	Low Diss	1	2.00	.000	3
		2	2.33	.577	3
		3	2.33	.577	3
		Total	2.22	.441	9
	High Diss	1	2.00	.000	3
		2	2.00	.000	3
		3	2.00	.000	3
		Total	2.00	.000	9
	Total	1	1.93	.594	15
		2	2.13	.516	15
		3	2.13	.640	15
		Total	2.07	.580	45
Total	Homog1	1	1.17	.389	12
		2	1.25	.452	12
		3	1.17	.389	12
		Total	1.19	.401	36
	Homog2	1	2.00	.000	12
		2	2.08	.289	12
		3	2.25	.452	12
		Total	2.11	.319	36
	Homog3	1	2.83	.718	12
		2	2.83	.718	12
		3	2.75	.622	12
		Total	2.81	.668	36
	Low Diss	1	2.08	.289	12
		2	2.33	.651	12
		3	2.17	.389	12
		Total	2.19	.467	36
	High Diss	1	1.92	.515	12
		2	1.92	.289	12
		3	1.75	.754	12
		Total	1.86	.543	36
	Total	1	2.00	.689	60

2	2.08	.720	60
3	2.02	.748	60
Total	2.03	.716	180

Note: Homog = Homogeneous Portfolio Level (1, 2, or 3), Low Diss = Low Dissonant Portfolio, High Diss = High Dissonant Portfolio