Running Head: Raters' Task Definition

Writing Assessment: Raters' Definition of the Rating Task

Mary L. DeRemer

Department of Educational and Counselling Psychology

McGill University, Montreal

A Thesis submitted in partial fulfillment of the requirements for the degree of

PhD in Educational Psychology

0-612-44407-4

Acknowledgments

I would like to express my appreciation to Bob Bracewell for the thoughtful guidance and sound advice which he gave as the supervisor of this dissertation. I value greatly what I have learned from him and want to thank him in particular for what he has taught me about writing. His patience, interest in this topic, and encouragement made a 'very significant difference'.

I would like to thank the three raters who participated in this study and the students who gave permission for their writing to be used. In addition, I would like to thank Muriel Fishman who transcribed the raters' think-aloud protocols, Tina Newman who did the reliability check of the coding method and Pascal Comeau who translated the Abstract.

My family and friends have helped me in important ways with their support and camaraderie, particularly P. Grafton, L. Senecal, J. Crammond, C. Kato, L. Papathanasopoulou, M. Jodoin, and E. Koritar and I thank them sincerely. I thank my mother for her prayers and my mother-in-law for her frequent trips to St. Joseph's Oratory to light candles for me. I would also like to thank H. Freedman and M. Zoccolillo for granting me educational leaves of absence from the MCH at critical times along the doctoral route.

Many thanks to dearest Stephanie, my most favorite student writer who has gone from scribbling to writing in two languages during the realization of this dissertation. She showed great patience while waiting for the "longest book in the world" to be finished.

I would like to express my greatest appreciation to my angelic husband, Mihai who has given me technical help and invaluable support during my doctoral studies. His love, amazing patience, humor, and willingness to forgive what the dissertation took from our family life means a great deal to me.

Finally, this study received financial support from Fonds pour la Formation de Chercheurs et l'Aide a la Recherche (FCAR), McGill University's Social Sciences Research Grant Subcommittee and the Laboratory of Applied Cognitive Science. I thank these sources for supporting this study and for the encouragement which the support afforded me.

Abstract

This descriptive case study examined how highly experienced raters do writing assessment, with a focus on how raters defined the task under two conditions: 1) as external raters and 2) as 'teacher as rater'. Three raters followed a think-aloud procedure as they evaluated student writing. The semantic structure of the think-aloud protocols was analyzed via the Task Independent Coding method. This analysis yielded a detailed representation of the objects and operations used by raters. The sequence which raters followed as they used these objects and operations was represented schematically by problem behavior graphs for each scoring decision made (N=360). Analyses of the problem behavior graphs showed that raters defined the task in three very different ways: 1) by searching the rubric to make a match between their response to the text and the language of the scoring rubric (search task definition), 2) by assigning a score directly based on a quick general impression (simple recognition task definition), or 3) by analyzing the criteria prior to score assignment without considering alternative scores (complex recognition task definition). Raters differed in their use of task definitions when they evaluated the same texts. These results challenged current writing assessment procedures which assume that raters internalize a scoring rubric during training and make a direct match between the scoring rubric and text characteristics. In addition, these results indicated that task definition is related to individual characteristics of the rater rather than status as a rater (i.e., external rater or 'teacher as rater'). These findings are discussed in terms of the effect of different task definitions on the validity of writing assessment.

Résumé

Cette étude de cas descriptive a analysé la façon dont trois évaluateurs expérimentés ont évalué des textes écrits par des étudiants. Plus spécifiquement, la manière qu'ils ont approché cette tâche est discutée selon deux conditions: 1) en tant qu'évaluateurs provenant de l'extérieur et 2) en tant que "professeur-évaluateur". La démarche d'évaluation des évaluateurs a été analysée suivant une méthode protocolaire de penser-à-haute-voix lors des évaluations des textes. La structure sémantique des protocoles obtenus par cette méthode protocolaire a été analysée selon la méthode de codification indépendante de tâche (Task Independent Coding). Suite à cette analyse, une représentation détaillée des objets et des opérations utilisés par les évaluateurs a été produite. L'enchaînement des objets et des opérations utilisés par les évaluateurs a été représentée schématiquement par des graphiques de procédures pour chaque décision (N=360). Les analyses de ces graphiques ont démontré que les trois évaluateurs définissaient la tâche de trois façons différentes: 1) en cherchant la rubrique du tableau référentiel d'analyse qui se rapprochait le plus de leur réaction face au texte (recherche de définition de tâche), 2) en assignant un bilan basé directement sur une impression générale sommaire (reconnaissance simple de définition de tâche), ou 3) en analysant les critères du tableau référentiel d'analyse avant d'attribuer un résultat, sans considérer d'autres possibilités (reconnaissance complexe de définition de tâche). Les évaluateurs ont utilisé différemment les définitions de tâche lorsqu'ils évaluaient les mêmes textes. Ces résultats remettent en question les procédures actuelles d'évaluation de texte qui assument que les évaluateurs internalisent les rubriques d'évaluation et qu'ils les mettent en relation directe avec les caractéristiques du texte. De plus, ces résultats indiquent que la définition de tâche est plus en fonction des caractéristiques individuelles de l'évaluateur que de son statut en tant qu'évaluateur (évaluateur provenant de l'extérieur ou "professeur-évaluateur"). Les effets qu'ont les différentes définitions de tâche sur la validité des évaluations de textes sont discutés.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

This study investigated the processes used by raters to evaluate writing. The issue of how a rater defined the rating activity in light of a standard psychometric approach to writing assessment was examined by analyses of the verbal protocols provided by raters as they evaluated student writing. The design of the study lent itself to the investigation of how a rater defined the rating activity as a) an external rater and b) as 'a teacher as rater'.

Rater Task Definition

To date there have been very few empirical studies which have investigated the processes used by raters to evaluate writing. This is surprising, given the importance and prevalence of large-scale writing assessment in the educational system. It is only through research which provides access to raters' verbalizations <u>during</u> the rating session that we can begin to understand how raters make judgments about writing quality. Results of studies which have used such a think-aloud methodology have shown that holistic raters adopted different rating strategies. Experienced holistic raters focused on different essay elements and had individual approaches to rating essays (Vaughan, 1991). Experienced raters made more comments after reading the text than did the inexperienced raters (Huot, 1993; Pula & Huot, 1993; Wolfe & Ranney, 1996). Wolfe and Feltovitch (1994) identified content focus and processing actions categories used by raters. The content focus of raters included appearance, the assignment, mechanics, organization, story telling, style and general. Processing actions included diagnose, monitor, review or rationale. They mentioned rater characteristics at a very general level and focused instead on a comparison of the content focus and processing actions used by "better" raters and "poorer" raters. They concluded that better raters stopped more often while reading essays to comment. However, these results were inconsistent with those reported by Huot (1993) and Huot and Pula (1993) and were not replicated in a later study by Wolfe and Ranney (1996) in which they found the following: First, raters at all levels of proficiency focused on similar text features. Second, while more proficient raters seemed to read a text without interruption and then evaluate it (i.e., interpret-then-evaluate), less proficient raters seemed to go through an alternating cycle of reading and evaluating portions of the text (i.e., interpret/evaluate/interpret/evaluate). Third, there was less variability between proficient raters' use of processing actions than there was between intermediate and non-proficient raters. Wolfe (1997) reported that less proficient raters who adopted a read/evaluate/read/evaluate strategy made evaluative decisions earlier and more frequently than did proficient raters. Nevertheless, these studies have failed to identify *how* proficient raters evaluate a text after reading it without interruption.

Scoring rubrics which identify criteria for assigning scores are relied upon for the achievement of reliable scoring. According to White (1984), the goal of rater training sessions is to help raters internalize the scoring rubric by combining description (the rubric) with example (the anchor texts). Well-trained raters score accurately and quickly and need only occasional reference to the rubric or anchor texts (p. 404). The assumption is that the criteria are sufficiently specific to enable consistency across raters in categorizing aspects of a piece of writing such as purpose, organization, details, etc. It is expected that following training raters will read a student's text or collection of texts and make a quick match between the rubric's criteria and the piece of writing. For example, it is estimated that it takes one to two minutes to rate a text holistically and one to two minutes to rate each criterion when a text is rated analytically (Spandel & Stiggens, 1980). Results of think-aloud research cited above tend to indicate that raters do *not* internalize the scoring rubric and make a direct match between the scoring rubric and text characteristics as they are apparently trained to do. That is, the use of processing actions as described by Wolfe and Feltovitch (1994) and Wolfe and Ranney (1996) and the nature and extent of the comments made by raters as revealed by the work of Vaughan (1991), Huot (1993) and Huot and Pula (1993) show that raters are involved in an activity which is more complex than a direct matching activity.

Status as Rater and Task Definition

A predominant feature of a psychometric approach to assessment is *independent* judgments by raters, yet the question of *who* should assess student writing has received little research attention. Given the increasing call for contextualized rather than decontextualized assessment of writing (Camp, 1993; Moss, Beck, Ebbs, Matson, Muchmore, Steele, Taylor & Herter, 1992; Witte, Flach, Greenwood and Wilson, 1995), it is important to know more about the rating processes of 'teachers as raters' and external raters. Pilot research reported by DeRemer and Bracewell (1995) indicated that 'teachers as raters' tended to see student texts as final drafts while external raters tended to consider the extent of semantic level revision needed and these differences may explain why external raters assigned lower scores on certain scoring criteria than did 'teachers as raters'. Koretz, McCaffrey, Klein, Bell, & Stecher, (1992) reported that on average, teachers did not rate their own students' writing portfolios more positively than did volunteer teacher-raters, but the Ministère de l'Èducation du Québec (MEQ) (1990) reported a study in which classroom teachers assigned scores higher than MEQ raters 43% of the time and scores lower than MEQ raters 5% of the time. (Teachers and MEQ raters agreed 52% of the time.)

The objective of this study then was to extend the results of previous think-aloud research in the area of writing assessment by identifying how highly experienced raters defined the writing assessment task. By investigating how a rater defined the assessment task, this research examined the meaning of the scores assigned. A second objective of this study was to investigate the task definitions constructed by a) pairs of external raters and b)'teachers as raters'.

CHAPTER II

Review of the Literature

There are two approaches to writing assessment: the traditional psychometric approach and an interpretative approach. Moss (1994) stated that in a typical psychometric approach each performance is scored independently by readers who have no additional knowledge about the student or about the judgments of other readers. Inferences about achievement, competence or growth are decontextualized, based on independent observations across readers and performances. The inferences are then referenced to relevant criteria or norm groups. Thus, the psychometric approach represents a standardized assessment and places emphasis on quantifying and rank ordering student's writing skills. In contrast, the interpretative approach involves collaborative inquiry that encourages challenges and revisions to initial interpretations. An interpretation might be warranted by criteria like a reader's extensive knowledge of the learning context; multiple and varied sources of evidence; and the transparency of the trail of evidence leading to the interpretations (Moss, p. 7).

Writing assessment practice historically has followed the psychometric tradition via direct and indirect formats. There are two main methods used to measure writing ability *directly* in large-scale assessment: 1) the assessment of an impromptu single writing sample and 2) the assessment of a collection of student writing (writing portfolio assessment). Writing ability is assessed *indirectly* through multiple choice tests which measure knowledge of standard written English and require no writing at all. This method involves machine scoring and does not involve human judgment of writing ability.

The literature review will be divided into psychometric and interpretative approached to writing assessment. First, research related to the validity and reliability of direct and indirect writing assessment formats in a psychometric approach will be reviewed. Second, research related to the validity and reliability of these same assessment formats in an interpretative approach will be reviewed. However, as a preface to understanding the validity and reliability issues which exist in writing assessment practice, current views of writing will be presented first.

Current Views of Writing

To measure growth and achievement in writing one needs a comprehensive understanding of writing. The current challenge for those who study writing and its development is to integrate social, cultural, and material factors that bear on writing with cognitive factors that underlie planning, writing, and revising text (Bracewell & Witte, 1997). Writing is social in the sense that the processes of reading and writing are always situated in particular social contexts and the meanings are constrained by what meanings are possible within and supported by those contexts. Readers and writers collaborate with other readers and writers because every new text is in some sense a response to at least one other text, which is in itself in response to at least one other text, and so on (Witte & Flach, 1994, p. 222). Bracewell and Witte (1997) provided the following account of the material and cultural aspects of writing. A writer communicates using material objects (letters, pens, paper, word processor, etc.) which in turn shapes the writing. The text which is the product of writing is a material object. The text also influences events in the material world. Writing is cultural in that cultural effects are part of a dialectic in which the individual characteristics interact with cultural characteristics to influence writing. In addition, "although writers rarely consider 'cultural' factors in an explicit manner, they certainly consider characteristics of their intended readership, and publicly honored characteristics of language (e.g., genre and register) that indicate an awareness of cultural constraints" (p. 4).

Writing is cognitive in that it is a problem-solving activity which draws upon the writer's memory, attention, knowledge, as well as factors related to problem representation, planning and idea generation (Hayes & Flower, 1980). Furthermore, writing is cognitive because writers often learn as they write. Engaging in symbolization processes such as reading and writing not only appears to mediate all learning, but would also appear (given people's memories of communication events) to insure that learning of some kind occurs when one engages in a meaning-constructive use of symbols (Witte & Flach, 1994, p. 222).

In any problem-solving activity the problem solver must represent the problem to him or herself, that is, understand the nature of the problem. However, problem representation in writing is a particularly complex process due to the ill-structured nature of the writing task (that is, there is no ready-made representation of the task and no standard solution procedure). The writer not only builds his or her own representation of the problem and its goals, but the problem or task itself changes as the constructed product grows (Flower, Schriver, Carey, Haas, & Hayes, 1989). For example, without concurrent feedback from an audience, the writer must anticipate the response of the audience as it reads the text. Consequently, task definition evolves during writing, with goals and subgoals changing as a result of evaluations of possible reactions to the emerging text (Bracewell & Breuleux, 1990). That writing is an ill-structured task has important implications for the learning which occurs during writing. To quote Bracewell and Witte (1997) "because one must elaborate the goal of an ill-structured task, the task context, which also includes one's current knowledge, necessarily changes in the course of doing it--these changes occurring because of the dialectic that occurs between one's knowledge and the evolving task definition" (p. 17).

Current cognitive models of discourse consist of levels of discourse representation and of processes that mediate these levels. The discourse structure of a text is characterized at different levels of representation, particularly semantic, surface structure and pragmatic levels. Theories and models of text production draw on this characterization of text discourse structure. For example, in the Frederiksen, Bracewell, Breuleux, and Renaud (1990) stratified model of text production, the production process proceeds from the specification of conceptual representation to the generation of sentences in a discourse. The writer must gradually constrain the production of semantic and linguistic structures. This is accomplished by constructing different levels of discourse representation and manipulating the fit among these representations so as to achieve a coherent discourse structure (Bracewell, 1987).

Validity and Writing Assessment

There are three traditional categories of validity evidence--content-related, criterion-related (predictive and concurrent) and construct-related--that operationally define validity at the present time (Standards for Educational and Psychological Testing-AERA, APA, NCME, 1985). However, the Standards are being revised (Linn, 1994) and there is growing consensus about the centrality of construct validity and the importance of expanding the concept of validity to include explicit consideration of the consequences of assessment use (Moss, 1992). Messick (1989) advocated two facets of validity specific to the consequences of assessment use: 1) the outcome of testing, and 2) the justification for testing. He distinguished the evidential basis of test use (evidence supporting the trustworthiness of score meaning) from the evidential basis of test interpretation (specific evidence for the relevance of the scores to the purpose of scoring and for the utility of the scores). He also distinguished the consequential basis of test use (appraisal of the value implications of score meaning) from the consequential basis of interpretation (appraisal of potential and actual social consequences of the testing).

There are radical changes taking place in educational assessment with a shift toward performance-based assessments (Linn, 1994). All writing assessments which yield a writing sample are considered to be performance assessments. However, not all performance assessments are considered to be authentic assessments. Meyer (1992) provided definitions which clarify the distinction between the two terms:

> In a performance assessment the student demonstrates the same behavior that the assessor desires to measure. If the behavior to be measured is writing, the student writes. In an authentic assessment the

> student not only completes or demonstrates the desired behavior but also does it in a real-life context. The significant criterion for the authenticity of a writing assessment might be that the locus of control rests with the student; that is, the student determines the topic, the time allocated, the pacing, and the conditions under which the writing sample is generated (p. 93).

Moss (1994a) discussed the tension between the disciplines of educational measurement and literacy education concerning writing portfolio assessment.

> Experience suggests that in order to achieve the standards of validity necessary for informing consequential decisions about individuals and programs, assessments need to be standardized to some degree. Standardization refers to the extent to which tasks, working conditions, and scoring criteria are similar for all students. Emerging views of literacy, however, suggest the need for less standardized forms of assessment to support and document purposeful, collaborative work by students (p. 110).

Alternate validity requirements have been suggested for performance assessments. Frederiksen and Collins (1989) proposed principles for the design of systemically valid testing which includes validity standards such as directness, scope, reliability and transparency. Linn, Baker and Dunbar's (1991) validation criteria include consequences, fairness, transfer and generalizability, cognitive complexity, content quality, content coverage, meaningfulness, and cost and efficiency.

Messick (1994) stressed that performance assessments must be evaluated by the same validity criteria, both evidential and consequential, as are other assessments. He recommended that where possible a construct-driven rather than a task-driven approach to performance assessment should be adopted because the meaning of the construct guides the development of scoring criteria and rubrics. He emphasized that focusing on constructs also makes salient the issues of *construct underrepresentation* and *construct-irrelevant variance*, which are the two main threats to validity (p. 14). As stated by Messick, the validity standard implicit in authenticity of assessment is minimal construct underrepresentation and the validity standard implicit in directness of assessment is minimal construct-irrelevant variance. Together they signal the need for convergent and discriminant evidence that the test is neither unduly narrow because of missing construct variance nor unduly broad because of added method variance (p. 22).

Reliability and Writing Assessment

Reliability is defined as the degree to which test scores are free from errors of measurement (AERA, APA, NCME, 1985, p. 19). The main sources of errors in the assessment of writing are the student, the test, and the scoring of the test or any combination of the above (Huot, 1990). Moss (1994) stated that typically, reliability is operationalized by examining consistency, quantitatively defined, among independent observations that are intended as interchangeable- consistency among independent evaluations or readings of a performance (i.e., reader reliability) and consistency among performances in response to independent tasks (i.e., task or "score reliability"). She noted that reliability is an aspect of construct validity (consonance among multiple lines of evidence supporting the intended interpretation over alternative interpretations).

Validity Issues in Psychometric Writing Assessment

Writing assessment methods yield scores and from the scores assigned, inferences are made about the growth and achievement of a writer. However, many factors may threaten the validity claimed for inferences of growth and achievement in writing. These factors include at least three components of writing assessment: a) the writing assessment method itself (Camp, 1993; Greenberg, 1992; Moss, Beck, Ebbs, Matson, Muchmore, Steele, Taylor & Herter, 1992; Witte, 1989); b) the scoring procedures used (Charney, 1984; Elbow & Blake Yancey, 1994; Moss et al., 1992); and c) the scoring criteria (Gere, 1980; Wiggins, 1994). These factors are discussed in turn below within the context of a psychometric assessment, that is, an assessment made by two independent judges.

Assessment Methods and the Validity of Psychometric Writing Assessment

Criterion-related evidence. Most of the research on the validity of writing tests has focused on criterion-related evidence, not construct related evidence (Greenberg, 1992). That multiple-choice tests show criterion-related evidence of validity is demonstrated by correlations between scores on multiple choice tests and performance on single writing samples (Breland & Gaynor, 1979; Godshalk, Swineford, & Coffman, 1966). Likewise, that impromptu essay tests show criterion-related evidence of validity is demonstrated by correlations between course grades and performances on impromptu essays (Breland et al., 1979; Godshalk et al., 1966). However, scores assigned to writing portfolios correlated poorly to moderately with classroom grades for writing (e.g., .29 to

.46, LeMahieu, Gitomer, & Eresh, 1995). In another study, substantial differences were found in students' performance when writing ability was judged based on a standard writing assessment, on individual samples of student work, and on portfolio collections as a whole (Herman, Gearhart, & Baker, 1993). Thus, criterion-related evidence has been demonstrated in multiple choice tests and single sample writing assessment but not portfolio assessment.

Construct-related evidence. Camp (1993) stated that multiple-choice tests do not sample the full range of knowledge and skills involved in writing nor do they sample writing skills in a manner which is consistent with theoretical constructs of writing. She related that this writing assessment method eliminates collaborative exploration and problem-solving by cutting off performance in writing from social and communicative contexts. Thus, it appears that these factors contribute to the construct underrepresentation of multiple-choice tests of writing ability.

The construct-related validity of the impromptu essay writing test has been questioned because it rarely provides an opportunity for students to engage in much of the process of writing, especially the rethinking and revising typical of the way experienced writers work (Moss et al., 1992). With the absence of opportunity for collaborative exploration in impromptu essay writing, one's performance in writing is also cut off from social and communicative contexts. This loss of communicative purpose and context is likely to be most damaging for students who are relatively unfamiliar with the mainstream culture (Camp, 1993, p. 57).

Research has not supported the construct-related validity of the impromptu essay test of writing ability. For example, Witte (1988) used evidence from think-aloud protocols to show the following: 1) that writers use different processes when writing in response to different tasks and 2) that differences across prompts can be attributed to the different demands that the prompts make on the writers' knowledge of the respective topics. These results demonstrated that it is unlikely that one can obtain a valid measure of writing ability based on evaluation of an impromptu single sample of writing.

Moreover, Witte, Flach, Greenwood and Wilson (1995) maintain that the impromptu essay test is decontextualized assessment. They stated that large-scale assessments like the NAEP (National Assessment of Educational Progress) use evaluation procedures "which are separated from naturally occurring language uses and purposes, and thus impose unnatural constraints on performance such that the performances themselves may become unnatural (i.e., artifacts of assessment)" (p. 61). They maintained that these evaluation procedures call into question the degree to which results of such assessments are actually indicative of underlying ability(ies).

Portfolio approaches to writing assessment appear to have the potential to accommodate the new constructs for writing (Camp, 1993). For example, features of a portfolio include the following: 1) multiple samples of writing gathered over a number of occasions; 2) variety in the kind of writing or purposes for writing that are represented; 3) evidence of process in the creation of one or more pieces of writing, and 4) evidence of reflection on individual pieces of writing and/or changes observable over time (Camp & Levine, 1990, p. 197). However, it is unclear whether the *procedures* used to evaluate portfolios yield a valid measure of writing ability. That is, the content of writing portfolios may represent writing constructs as they are currently understood, but the procedures and criteria used to evaluate the portfolios may not capture all or even part of the writing constructs contained within the portfolio. This issue is discussed at length below.

Scoring Procedures and the Validity of Psychometric Writing Assessment

The procedures used to score multiple-choice tests are not a validity issue because these tests are considered to be "objective" (i.e., the answer is either correct or incorrect) and are usually scored by a machine. In contrast, when a sample(s) of writing is evaluated, the evaluation is considered to be "subjective" (i.e., determined by and emphasizing the ideas, thoughts and feelings of the rater).

There are three main scoring procedures used by raters to evaluate the quality of writing samples when an impromptu essay or a writing portfolio is assessed : primary trait, holistic, and analytic (Huot, 1990a). In primary trait assessment, the rhetorical situation creates the criteria for the evaluation and a scoring guide which is specific to the genre of the writing task is developed for each task (Lloyd-Jones, 1977). When a rater uses a holistic procedure the rater assigns a single score to a text or a set of texts. Typically the assigned score subsumes performance on multiple criteria such as purpose, organization, grammar, etc. and represents a value on a continuum which ranges usually from one (the lowest score) to four or six (the highest score). For example, to be assigned a score of three on a holistic scale used in the Huot (1993) study the text had to meet the following criteria: shifting focus; shifting tone; less clear development; and minor surface problems.

When a rater uses an analytic scoring procedure he or she assigns multiple scores to a text, one score per scoring criteria specified in the rubric such as purpose, organization, details, voice and tone, grammar, usage and mechanics as seen in the Vermont Writing Assessment Program (Koretz et al., 1992). These scores also represent a value along a continuum. As noted above, raters are trained in the use of these procedures and the assumption underlying such training procedures is that raters will internalize the scoring rubric.

Holistic scoring and the impromptu writing sample. Holistic scoring procedures are the most widely used writing assessment procedures (Huot, 1990a). The validity of using holistic scoring to assess writing (i.e., single sample) has been questioned in the writing assessment literature (i.e., Charney, 1984; Elbow & Blake Yancey, 1994; Gere, 1980) yet there has been very little research which has investigated this question. Charney (1984) stated that in order to achieve a high reliability, testing agencies and researchers must impose a very unnatural reading environment, one which intentionally disallows thoughtful response to the essays. She identified the speed at which raters are recommended to work (e.g., one 400 word essay per minute), the peer pressure to conform to a given set of rating criteria, and the frequent monitoring during rating sessions as disruptive to the reading process. In addition, she stated that rating criteria have only *ad hoc* validity; they may be acceptable only to the group which formulates them.

Huot (1993) used a think-aloud procedure to investigate Charney's (1984) objection that holistic ratings are generated by scoring procedures which alter fluent reading processes and impede the quality of raters to make sound judgments of writing quality. He compared the rating processes of four experienced holistic raters who received training and used a scoring rubric with the rating processes of four inexperienced raters who did not receive training and who did not use a scoring rubric. Raters evaluated individual texts. Results of this comparisons indicated that experienced raters made more personal comments than the inexperienced raters and they contributed a wider variety of responses than the inexperienced group. Huot concluded that holistic rating did not impede true and accurate reading and suggested that holistic scoring procedures actually promote the kind of rating process which insures a valid reading and rating of student writing. Thus, while Huot concluded that the results suggested a valid reading and rating of student writing, he also stated that these results cannot be used to infer construct validity for holistic scoring.

Holistic assessment and writing portfolio assessment. Researchers have begun to investigate procedures for assessing writing portfolios, yet the question of how one should evaluate a writing portfolio remains unanswered (Calfee, 1994a). In particular, a key question is whether a score for a writing portfolio should be derived from judgment of the portfolio in totality (i.e., holistically) or from the sum of its individual pieces (Baker & Linn, 1992). Those who favor holistic assessment of a writing portfolio call for the rater to hold his or her judgment in abeyance not only over the course of a single essay but over the course of an entire portfolio (Sommers, Black, Daiker, & Stygall, 1993). Others maintain that readers are bound to consider the multiple texts in light of one another, weighing their strengths and weaknesses and finally reaching a single judgment based on the parts, not a dominant impression of the whole portfolio (Hamp-Lyons & Condon, 1993). Moss et al. (1992) stated that growth is often manifest in qualitative changes in the writing which involves comparisons of student's revisions in multiple texts. They maintained that to average the scores from multiple scores so as to talk about achievement in writing or to subtract or otherwise manipulate the scores to talk about growth would miss the point (p. 13).

However, the current practice in writing portfolio assessment is for raters to assign one score to the portfolio (Allen, 1995; Condon & Hamp-Lyons, 1994; Sommers et al., 1993). In some cases the portfolios is assigned one score for each of several dimensions or criteria. For example, writing portfolios in the Pittsburgh Public Schools are assigned one score on each of the following three dimensions: accomplishment as a writer, use of processes and resources, and growth and engagement as a writer (LeMahieu et al., 1995). Writing portfolios in the Vermont Assessment Program are assigned one score for each of five criteria: purpose, organization, details, voice and tone, and grammar, usage, and mechanics (Koretz et al., 1992).

There has been minimal investigation of the validity of the scores assigned to writing portfolios. Research by Nystrand, Cohen, and Dowling (1993) indicated that the writing ability of the students was not consistent across the different genres contained in the portfolio. They concluded that the strategy of characterizing the contents of a portfolio with a single writing ability score failed to reflect the heterogeneity of the texts.

Purves (1992) drew the same conclusion when he reported on a study of achievement in written composition that involved students, teachers, and researchers in fourteen countries. All students wrote in response to three tasks. Researchers found strong independence among the task scores, sufficient independence to prevent summing them into some construct such as "writing performance" or "writing ability" (p. 5).

Gearhart, Herman, Baker and Whittaker (1992) reported that raters indicated that they felt that the mix of genres in a portfolio obscured evidence of the components of the writing process and evidence of changes over time in writing quality. As a result raters were able to assign only a General Competence score to the portfolios. Gearhart et al. concluded that it is possible to score portfolios consistently- if the aim is to reduce them to a single score of overall quality; however, most importantly, they concluded that the results of the study raised serious questions about the meaningfulness of the scores assigned to the portfolios. Further uncertainty about meaning of a single score, that is the validity of a single score, was expressed by Dickson (cited in Allen, 1995) who stated that raters who assessed writing portfolios agreed on a final judgment but for different reasons.

The implications of research which has investigated the construct validity of assigning a single holistic score to an impromptu essay or assigning a single holistic score to a writing portfolio are clear. In the first case (impromptu essay), given the variability in writing skill across task and the different processes used to write in response to different tasks (Witte, 1988), it is unlikely the score assigned to an impromptu essay can represent 'writing ability' but it may represent the writer's ability to write in response to that specific task. Thus, research by Witte (1988) revealed the construct underrepresentation of writing assessment based on an impromptu essay. Furthermore, it is argued that writing an impromptu essay necessitates the use of processes other than those which represent writing as it is now understood (Witte et al., 1995), thus presenting an instance of construct irrelevant variance. Thus, research has demonstrated both the construct underrepresentation and construct irrelevant variance of assessment methods using the impromptu essay.

In the second case (writing portfolio assessment) research has shown that the current practice of assigning a single score to the portfolio may fail to capture the variability of the scores assigned to the different texts within the portfolio (Nystrand et al., 1993; Purves, 1992). As such, the single score holds questionable meaning beyond representing an average of performances on different tasks. This uncertainty about the meaning of the score poses a serious threat to the validity of inferences drawn from a holistic score assigned to a writing portfolio.

Scoring Criteria Used and the Validity of Psychometric Writing Assessment

Greenberg (1992) stated that the question of substantive criteria for "good writing" relates directly to the issue of construct validity. She maintained that the skills described in the criteria on current holistic scoring guides do not provide an adequate definition of "good writing" or the many factors that contribute to effective writing in different contexts. Gere (1980) maintained that existing systems for investigating writing (i.e., holistic, analytic and primary trait evaluation) share the common weakness of ignoring the communication function of meaning. She wrote that primary trait evaluation appears to accommodate communication intention but does not provide for genuine communication intention because it limits the kind of meanings the reader can consider (p. 48). Wiggins (1994) stated that writing rubrics in every district and state over-emphasize formal, format, or superficial trait characteristics (p. 132). In addition, Myers and Pearson (1996) maintain that approaches used to score one writing genre (e.g., an editorial) cannot be the same as those used to score another (a report or an autobiography) (p. 14). However, it is standard practice for a single rubric to be used to assess a portfolio which includes a variety of genres written in response to different tasks. Furthermore, it has been questioned if raters actually apply the criteria they have been trained to use (Charney, 1984). Thus, there is the possibility of both construct underrepresentation and construct irrelevant-variance with the scoring criteria presently used.

The literature on what criteria raters chose in judging writing quality can be divided into two types: *correlational research* and *think-aloud research*. Correlational research focuses on a) the correlation between textual features and quality scores, and b) the correlation between the general aspects of quality scores and content, organization, and mechanics (Huot, 1990b, p. 206). Results of the correlational research indicated that raters are mostly concerned with content and organization (see Huot, 1990a for a comprehensive review of the literature of direct writing assessment).

Results of think-aloud research have been consistent with results of earlier correlational research. As reported above both inexperienced and experienced raters in the Huot (1993) study made more comments about the content and the organization of the text than any other criteria. The same results were found when the study was replicated by Huot and Pula (1993). Vaughan (1992) reported that raters made the most comments about content and handwriting as well as criteria that were not present in the scoring rubric used in the study. Wolfe and Feltovich (1995) analyzed the think-aloud protocols of six raters and found that the raters cited development, organization, and voice most often as considerations for scoring. These three aspects were given the most emphasis on the scoring guide. Wolfe and Ranney (1996) also found that regardless of level of inter-rater agreement, scorers focus on similar features of an essay as they formulate scoring decisions using a narrative scoring rubric. Raters focused with the greatest frequency on the criteria storytelling (ability to tell a story) and organization.

Thus, research demonstrated consistently that raters focus most frequently on the scoring criteria content and organization. These results provide evidence that there is a poor fit between the scope of *what* is measured in writing assessment and current writing constructs. For example, a better fit between scoring criteria and writing constructs might include asking the following questions as suggested by Wiggins (1994): Can students make good use of feedback, can students profit from self-reflection, are they developing a better grasp of what does and does not work, and are they getting better at judging the value of the feedback they receive (p. 138). Clearly, the scoring criteria used to assess single texts and writing portfolios appear to reflect construct underrepresentation, which as stated earlier is one of the major threats to validity.

Gearhart and Wolf (1994) in response to earlier research which showed that teachers constructed a set of criteria to guide writing assessment that made no reference to genre and emphasized mechanics and generalized features of writing content, designed a training study to enhance teachers' knowledge of narrative text and teachers' competence with methods of narrative assessment. They emphasized an understanding of the components of the narrative (i.e.,: genre, theme, character, setting, plot, style, tone, and point of view) and the technical language that represent narrative content. To encourage teachers to offer explicit guidance for their writing they developed a narrative feedback form for written commentary and a narrative rubric for judging the effectiveness of students' narratives.

They reported that prior to training teachers rarely characterized narrative writing with a technical language that captured its heart or complexity. Following training all teachers reported perceived change in their understanding of narrative. However, seven months later, questionnaires and classroom observations indicated that weaknesses in teachers' understanding of narrative continued to affect their methods of narrative assessment. In training sessions most teachers demonstrated a capacity to understand and use the Writing What You Read (WWYR) rubric and feedback form effectively . However, in the classroom teachers rarely used the narrative feedback form for written commentary or the narrative rubric for scoring. Instead, teachers used the narrative feedback form and the rubric to design assignments, establish criteria, and assess narratives "even if the assessments were oversimplifications of the rubric's components".

In summary, both multiple-choice tests of writing and single sample writing assessment have shown criterion-related evidence of validity but not construct-related evidence of validity. Writing portfolio assessment appears to have the potential to met construct-related validity requirements but it appears that the main procedure used to assess writing portfolios (i.e., holistic scoring) simultaneous with the criteria incorporated in holistic scoring procedures seriously undermines this potential to address fully construct-related validity requirements.

Reliability Issues in Psychometric Writing Assessment

The multiple-choice test, with its machine scoreable items, has been seen as reliable (Camp, 1993). The reliability of scores assigned to a single writing sample is considered to be high when texts are rated in well-controlled rating sessions as described earlier (Camp, 1993; Charney, 1984). However, texts cannot be rated reliably without the use of rigorous training procedures (Witte, 1993) and sometimes reliable rating is not achieved even with rigorous training and a controlled testing environment. For example, the Ministère de l'Èducation du Québec (MEQ) (1990) reported that independent raters disagreed with each other 25% of the time.

Research has shown that it is possible to assess writing portfolios consistently if the aim is to assign a single score to the entire portfolio (Baker & Linn, 1992). High reliability figures were reported when portfolios were assigned a single score (Allen, 1995; Gearhart et al., 1992; Sommers et al., 1993) and a single score on three dimensions (LeMahieu et al., 1995), yet the validity problems associated with this have been discussed above. However, rater agreement was low when raters assigned scores to multiple scoring criteria (Koretz et al., 1992; Resnick & Resnick, 1993).

While high rater reliability figures are often reported in the literature, the reliability data reported for single text assessment and writing portfolio assessment is often ambiguous and easy to misinterpret. This is because reliability can mean at least two things. First, it may mean that raters assign the same scores to each text or portfolio, or second, it may mean that a high correlation was obtained between scores assigned by raters. The first logically implies the second but the reverse does not hold true. Further a correspondence of average scores implies neither. Statistically, it is possible to have cases where one rater assigns higher scores and one rater assigns lower scores with the result that scores assigned to texts differ between raters but the correlation between scores is fairly high. Consequently, it is difficult to determine on the basis of reliability coefficients or correspondence of average scores if raters are in fact assigning the same ratings to individual texts.

Cherry and Meyer (1993) have also identified several important problems with reliability in writing evaluation. First, they noted that discussions of reliability have typically been limited to inter-rater reliability thus excluding discussion of instrument reliability. They stated that whereas inter-rater reliability describes how consistently raters judge the *writing quality* of writing samples, instrument reliability addresses the reliability of judgments of *writing ability* made on the basis of those samples (p. 114). They maintain that by way of describing how consistently an assessment instrument measures the performance of a particular group of students on a particular kind of writing task scored in a particular way, instrument reliability comes close to describing how valid the assessment is within the given constraints. Second, they noted a lack of agreement on appropriate statistics for calculating and reporting inter-rater agreement. For example, some studies report per cent agreement figures which represent 100% agreement and other studies report figures which represent agreement within one point on a four to six point scale. Third, they noted that the standard practice of resolving differences between two raters by seeking a third rating is a serious problem. Thus, when raters disagree by more than one point, usually a third rating is obtained and the "bad" rating of the three is thrown out. Interrater reliabilities are calculated on the basis of the new set of paired ratings. However, the resulting coefficient will be both inflated and largely meaningless (p. 123).

Thus, it is very difficult to know on the basis of the reliability coefficients and per cent agreement figures reported in the literature whether or not the raters assigned the same scores to the same texts or portfolios, that is, whether the assessment yielded consistent judgment of a student's writing ability.

Validity Issues and an Interpretative Approach to Writing Assessment

In the previous section factors which may threaten the validity claimed for inferences of writing ability in a psychometric approach to writing assessment were discussed. These factors included the writing assessment method itself, the scoring procedures used and the scoring criteria developed for the assessment. In the section below, these same factors will be discussed within the context of an interpretative approach to writing assessment.

Assessment Methods and the Validity of an Interpretative Approach to Writing Assessment

Indirect writing assessment methods (i.e., machine scored multiple choice testing) are the antithesis of interpretative assessment and are not discussed here. Evaluation of the impromptu single essay appears to preclude an interpretative approach to assessment. This is because when one assesses an impromptu single sample only, one can evaluate neither revision of a text nor evaluate revision across texts-- evaluations which are an integral part of interpretative assessment of writing (Moss et al., 1992).

Writing portfolios are viewed as valuable in an assessment model in which teachers' interpretations of their students' growth and achievement play a central role (Moss et al., 1992). Given that an inductive approach to writing assessment is only warranted within the context of a writing portfolio, the following sections on the validity of an interpretative approach will concern writing portfolio assessment only.

Scoring Procedures and the Validity of an Interpretative Approach to Writing Assessment

Moss et al. (1992) provided the following example of an interpretative assessment of a writing portfolio. First, they developed a list of features to be used in analyzing the contents of each piece contained in the portfolio such as the plans, drafts, final draft, student's self-reflections, the teacher's reflections, etc. This list of features comprised a framework which was an intermediate step undertaken to inform the writing of the narrative profiles which described the student's achievements and growth in writing. They maintained that the narratives taken together with the frameworks and the portfolios allow another reader to serve as co-analyst, tracing the evidentiary trail that led to the conclusions and raising alternative interpretations for discussion. They viewed differences of opinion between readers as opportunities for discussion and rethinking of initial interpretations. They stated that if the approach described here is used as intended, the central interpretation will be that of the classroom teacher and it will be based not only on the portfolios but also on extensive knowledge of the student, their goals, and their instructional opportunities.

Moss et al. (1992) investigated the validity of portfolio-based interpretations by 1) investigating the validity of the ratings as reflected in the narrative profiles and 2) by investigating the representativeness of portfolio selections. They selected 10 students and examined both their writing folders (which contained all the writing during the year) and their portfolios (which contained the pieces which the students had selected to represent themselves as developing writers). Sets of raters independently wrote narratives based on the folders and the portfolios. A content analysis showed that there were substantial difference in emphases among readers in the written narratives. After reading the folders from which the portfolios were selected, the raters concluded that the students occasionally left out what the raters perceived to be the stronger pieces out of the portfolio. Some students gave in-depth information with respect to a particular genre but little information about other genres. Other students gave a broad sampling of writing across genres but insufficient samples to note changes within genre. The authors noted that the portfolio selection process is a complex problem and requires important decisions in order to balance student's autonomy with teachers' informational needs.

Moss (1996) provided an illustration of a proposed partial interpretative approach to evaluating portfolios in the context of teacher certification which has a clear application for writing portfolio assessment. This evaluation procedure appears to represent a shift away from the central role of one individual's interpretation (i.e., the teacher) to a collaborative effort. In the partial interpretative method which is being planned by Moss individual readers will first work through the portfolio alone, noting and recording evidence relevant to the interpretative categories which have been established for the assessment. Raters then will work together to prepare interpretative summaries with supportive evidence for each category. The performance standards will be operationalized through multiple exemplars of performance. After completing interpretative summaries and supporting evidence records, raters will debate and reach consensus on an overall level of performance. Then they will prepare a written justification tying the evidence they have analyzed to a decision. The decision, written justification, and interpretative summaries with supporting evidence will be audited by a criterion reader who may or may not recommend more extensive review. A sample of portfolios will be evaluated by a second pair of readers as part of the ongoing monitoring of the system (p. 25).

In the interpretative approach to teacher certification outlined by Delandshere and Petrosky (1992) raters are required to make the reasoning behind their judgments explicit by answering critical questions in the writing of interpretative summaries. For example, a question concerning learner-centeredness is stated as follows: How does the candidate anticipate students' abilities to interpret literature through discussion and accommodate students' thinking in the discussions, and are activities related to the discussion? (p. 14). Raters are also required to assign ratings on a scale from one to four based on the interpretative summary written for every dimension pertinent to a given task. Raters translate their interpretative summaries into numerical ratings by comparing the examined performance to decision guides developed to synthesize the differences in candidate's conceptual understanding and to represent the different points on the rating scale for all dimensions on each task (p. 15). A second rater reads and/or observes the original performance as well as the interpretative summaries and confirms or disconfirms the plausibility of the interpretations and the consistency with which the evidence of the performance leads to the interpretations and judgments.

Scoring Criteria and the Validity of an Interpretative Approach to Writing Assessment

In the interpretative approaches to validity described above (Moss et al., 1992; Moss, 1996) there is a shift away from specific scoring criteria which are associated directly with a score. Instead, raters use either a list of features, or interpretative categories which guide the writing of an interpretative summary. Taken together, the organization of the frameworks (i.e., list of features) and the three to five "sequences" within the portfolio (i.e., final draft; plus all related preliminary drafts or plans; self-reflections about reasons for selection, the strength of the writing, and goals for subsequent work; and teachers and other reflections about the writing) allow raters to look at consistencies among the features of the different "sequences" contained in the portfolio. Thus, it is possible to examine the extent to which students seem to be setting goals for themselves, using others' comments, and following through in revision and showing improvement in subsequent pieces of writing. Progress and achievement noted in these areas become integral parts of the interpretative summary (Moss et al., 1992, p. 18).

Reliability Issues and an Interpretative Approach to Writing Assessment

Moss (1994) stated that epistemological and ethical concerns about reliability concern the extent to which one can generalize the construct of interest from particular samples of behavior evaluated by particular raters and the extent to which those generalizations are fair. With respect to generalization across tasks, the goal of an interpretative approach is to construct a coherent interpretation of collected performances. Inconsistency in students' performance across tasks does not invalidate the assessment. Rather it becomes an empirical puzzle to be solved by searching for more comprehensive or elaborated interpretation that explains the inconsistency. With respect to generalization across readers, Moss (1994) stated that an interpretative approach to assessment privileges interpretations from readers most knowledgeable about the context of the assessment. Initial disagreements among raters would provide an impetus for dialogue, debate, and enriched understanding informed by multiple perspectives as interpretations are refined and as decisions or actions are justified. Thus, interpretative assessment activities serve the same purpose as multiple independent readings serve--warranting the validity and fairness of the approach (Moss et al., 1992).

Moss (1994) concluded that there can be validity without reliability when reliability is defined as consistency among independent measures intended as interchangeable. She stated that reliability serves important purposes such as indicating the extent to which we can generalize to the construct of interest from particular samples of behavior evaluated by particular readers and the extent to which these generalizations are fair. However, she maintains that an interpretative approach to assessment provides a means of serving those same purposes.

As explained above, from a conceptual viewpoint, it is difficult to determine on the basis of reliability coefficients or correspondence of average scores if raters are in fact assigning the same rating to individual texts. And, from a technical viewpoint, following the argument of Cherry and Meyer (1993) concerning the present problems in *reporting* reliability data, it is very difficult to know if adequate consistency among independent ratings truly is obtained in writing assessment research and practice. As noted above, rater consistency is most readily obtained when pass/fail judgments are made or a holistic score is assigned to a writing portfolio yet the attendant problems for the validity of these judgments is well understood. Thus, when the concept of reliability is viewed in light of the problems currently associated with it, then it becomes unclear if the power afforded to the concept is warranted in the domain of writing assessment--that is, must an assessment of writing be deemed "reliable" in order to be considered valid when the value of reliability data is questionable on conceptual and technical grounds?

Who Should Assess Student Writing

The question of who should assess student writing (the student's own teacher or an external rater) has implications for the validity and reliability of writing assessment in both psychometric and interpretative approaches. There appears to be a lack of consensus among researchers concerning this question. Moss et al. (1992) state that the use of narrative profiles in an interpretative approach acknowledges the singular value of the teacher's knowledge base in making interpretations which can not be duplicated by outside readers. Calfee (1994b) maintains that the classroom teacher is arguably in the best position to make informed judgments. Others such as Resnick and Resnick (1992) stated that using performance assessments as part of public accountability programs would require that students' performance is evaluated by panels of judges other than the student's own teacher. Mehrans (1992) stated that when assessing for accountability purposes, it is imperative to have performances scored by those who do not have a vested interest in the outcome. Having teachers score their own students' performances fails this principle (p. 8). As noted above, the limited research is equivocal on rater bias, with two studies reporting bias (DeRemer & Bracewell, 1995, MEQ, 1990), and another reporting none (Koretz et al., 1992).

Rationale

The writing of nearly every student in North America will at some point be assessed as part of a large-scale writing assessment program yet very little is known about a) *how* decisions are made about growth and achievement in writing, b) who should be assessing student writing, and c) the validity of these judgments of growth and achievement in writing. Writing assessment practice has been built on the assumption that during training raters internalize a scoring rubric which they apply directly to student texts. This study builds on research using think-aloud methods which challenged this assumption by showing that raters showed a high level of personal engagement with the texts which they evaluated (Vaughan, 1991; Huot, 1993; Huot & Pula, 1993) and that raters used specific processing actions (Wolfe & Feltovitch, 1994; Wolfe & Ranney, 1996; Wolfe, 1997). Given the discrepancy between assumptions underlying writing assessment practice and results of think-aloud studies, one objective of this research was to investigate how experienced raters defined the rating task. A second objective of this study was to investigate the task definitions constructed by a) pairs of external raters, and b)'teachers as raters'. By investigating how a rater defined the assessment task, this research examined the meaning of the scores assigned.

Contributions to Knowledge

The study presented below yields an original contribution to knowledge. First, unlike previous think-aloud research which used coding methods which were extracted from the think-aloud protocols themselves, in this research a theoretically motivated coding method was applied to the data. Second, by adopting a case study methodology with a small number of raters, it was possible to construct problem behavior graphs made up of the objects, operations, and relations which were identified by the theoretically motivated coding method. Problem behavior graphs have been used to understand the problem-solving activity of subjects in other domains but they have not been used to study the writing assessment process. These problem behavior graphs represented the *sequence* of rating activity followed by each rater for every text which was evaluated. They were instrumental in the identification of the task definitions constructed by raters. Third, the design of the study also yielded an understanding of the differences in task definition found between a) pairs of external raters and b) 'teachers as raters'. Previous research (think-aloud and non think-aloud) has focused on expert-novice comparisons rather than yielding a fine grained analysis of the behavior of individual raters.

Finally, the outcomes of this study bear on the theoretical domain of writing assessment. Writing assessment practice has progressed without a theory of writing assessment (Gere, 1980; Witte, 1988) and these results begin to show that any emerging theory of writing assessment should incorporate social and cognitive, material, and cultural factors as in a theory of writing (Bracewell & Witte, 1997).

CHAPTER III

Methods

A descriptive case study methodology was used in which the activity of the individual rater in assessing texts constituted the case. This methodology was used for a number of reasons.

First, in order to examine how raters define the activity of text evaluation for themselves, a methodology was needed which allows for a very fine grained level of analysis at the level of the individual rater. Second, given the theoretical orientation of this study which views writing and the evaluation of writing as social cognitive processes, it is assumed that the phenomenon of writing evaluation cannot be studied outside of its social context. Finally, the question of how a rater defined the activity of evaluating writing lent itself to investigation of multiple sources of evidence. In this research it was possible then to investigate the rating activity of the individual rater as well as the actual scores assigned by the rater.

However, the research methodology followed here departs from a descriptive methodology in one important way. In more traditional case studies, 'pattern coding' during data collection is central to the analysis (Miles & Huberman, 1984). Pattern coding is an inferential process which consists of reading the data collected to date to see what patterns emerge. The identified pattern is then coded and this pattern code is then tried out on the next set of transcribed field notes or documents to see if the pattern fits the new data. The most promising patterns are then written up in the form of a memo that provides support for the significance of the code. Finally, pattern codes are checked out in the next wave of data collection. In contrast, in this study an *a priori* theoretically-motivated coding method which is discussed below was applied to the data.

Participants

Three highly experienced raters participated in this research. All were grade eight English teachers who used the Vermont Writing Assessment Analytic Assessment Guide (1991) extensively for instruction and assessment purposes. They were Vermont Writing Network leaders who trained other English teachers throughout the state of Vermont in the use of this scoring rubric. These raters were chosen because of their extensive experience with the rubric and the leadership role they assumed in training other raters in the use of the rubric. In addition, given the increased prevalence of local scoring (as opposed to central scoring) in large scale assessment it was important to determine how classroom teachers evaluate student writing. The design of this study lent itself to a comparison of differences in task definition between external raters and teachers who rated their own students' writing. Two of the raters were the teachers of students whose texts were evaluated as part of this research. Rater 1 (Pat) had taught four of the students who provided texts for assessment (Set A), and Rater 2 (Tom) had taught the remaining four students who provided texts (Set B). Rater 3 (Kathy) had not taught any of the students and hence, acted as an external rater for all of the texts. These relations between rater and teaching status are depicted in Figure 1.

Raters were paid for their participation in this research. The research procedures described here were considered to be acceptable on ethical grounds by the Research Ethics Committee of the Faculty of Education, McGill University (see Appendix 1).

Materials

Raters used the scoring rubric of the Vermont portfolio assessment procedure (Vermont Department of Education, 1991).[1] This rubric consists of five scoring criteria: purpose, organization, details, voice and tone, and grammar, usage and mechanics (GUM). For each of the five criteria one of the following values on a four point scale was assigned: Extensively, Frequently, Sometimes, and Rarely. Operational definitions of quality at each point of the scale are provided for each scoring criterion. The scoring rubric used in this study is presented in Figure 2.

The first three texts from eight writing portfolios produced by grade 8 students were studied. The three texts were as follows: a Letter of Introduction, A Best Piece and a Letter about the Best Piece. Twenty four texts were rated by each of the three raters.

Procedure

Task Procedure

Raters met individually with the experimenter. They were informed that their task was to assess each text following the guidelines of the Vermont writing assessment program. They were also instructed to think aloud while they implemented the given criteria of assessment. That is, they were instructed to verbalize all their thoughts and impressions throughout their evaluation of each text. (See Appendix 2 for instructions given to raters). Raters practiced using the think-aloud method on two texts prior to

[1] A revised version of this rubric is currently used in the Vermont writing assessment program.

beginning the rating session. All twenty four texts were evaluated in the same order by each rater.[2]

Subjects' verbalizations were recorded on audiotape. The think-aloud protocols of each rater were transcribed and the transcriptions were segmented into clausal units. The number of clausal units totaled about 10, 000.

Analysis procedures

Rater Agreement

Interrater agreement (Pearson r and percent agreement) were calculated.

Analyses of Raters' Problem-solving Activity

When one evaluates student writing using a scoring rubric one is engaging in a problem-solving activity. Problem solving is defined here as a behavior directed toward achieving a goal (Anderson, 1990). The rater's goal is to make decisions about the quality of student writing based on a given set of guidelines which are applied to characteristics of student compositions. As part of this activity raters must abstract the set of guidelines written in the rubric. For example, a rater reading an excerpt from the scoring rubric's criteria for details- 'details lack elaboration' -must interpret the language of the rubric and

[2] Raters use the Vermont Writing Assessment-Analytic Assessment Guide in conjunction with benchmark texts when assessing student writing portfolios (G. Hewitt, personal communication, May 30, 1997). However, the raters in this study did not use benchmark texts because of their extensive experience with the scoring rubric and their background as Vermont Writing Network leaders. It is possible that the inter-rater agreement reported here may have been higher if benchmark texts had been used. However, the level of inter-rater agreement reported here is consistent with earlier agreement levels reported by Koretz et al. (1992) when benchmark texts were used by raters.

then reconcile this interpretation with the specifics of the text. Thus, evaluating writing when using a scoring rubric is a *constructive* activity. If this is so, contrary to assumptions inherent in the rater training process discussed earlier, the activity in which the rater engages *cannot* be considered a simple match between the specifics of the text and the criteria set out in the scoring rubric. Instead, how the rater represents and elaborates the rating activity can be analyzed from a problem-solving perspective.

In the literature on problem solving a distinction is often made between well-structured problems and ill-structured problems. However, Simon (1978) stated that there is no precise boundary between problems that may be regarded as well-structured and those that are ill-structured. The distinction describes a continuum and not a dichotomy. Simon identified three key features which distinguish ill-structured problems from well-structured problems. First, in ill-structured problems the criterion that determines whether the goal has been attained is both more complex and less definite. Second, the information needed to solve the problems is not entirely contained in the problem instructions and the boundaries of the relevant information are very vague. Third, there is no simple "legal move generator" for finding all of the alternative possibilities at each step (p. 286). Voss and Post (1988) stated that an important question for the solving of all ill-structured problems is that of what constitutes a good solution. They related that generally there are not “right answers” to ill-structured problems.

Writing assessment is an example of an ill-structured task. Despite standardized training procedures, there is no standard solution procedure for writing assessment. Typically raters are trained to agree with each other by reaching consensus on sets of anchor papers. Raters are given a scoring rubric and benchmark texts, not standard solution procedures. As such, they are presented with a task environment (i.e., assess the texts using a given rubric) but they are not given a ready-made representation of the rating activity. The rater must develop his or her own plan of action. The rater must also define those goals and criteria which will themselves represent the activity (e.g., What will constitute 'rudimentary development of ideas' in this situation ?).

Analysis of Raters' Construction of the Task

The analysis of the raters' construction of the assessment task followed the methods used by DeRemer and Bracewell (1991) to investigate the assessment activity of holistic raters. This methodology drew on three sources for determining the *knowledge* (i.e., objects) and *processes* (i.e., operations) used by raters. These sources are summarized in Figure 3.

Task analysis of the rating procedure. A task analysis (see Ericsson and Simon, 1993, p. xv) of the rating procedure yielded minimal information because, as stated earlier, a standard rating procedure does not exist. However, the task analysis did yield a set of goals contained in the task instructions (e.g., evaluate the sets of writing portfolios in the order in which they are presented to you using the given scoring rubric), a set of possible objects inferred from world knowledge (e.g., rubric or author) and a set of possible operators inferred from world knowledge (e.g., select a particular text to read, reread the rubric).

Analysis of the nonverbal activity of raters. This analysis yielded such operations as 'reading text' or 'reading the scoring rubric'.

Analyses of think aloud verbalizations. The analyses of the think-aloud verbalizations were based on the Task Independent Coding methodology developed by Bracewell and Breuleux (1994). When applied to the rating task this methodology yielded a detailed representation of the objects and the operations that the raters used in evaluating a text. Task Independent Coding calls for the think-aloud protocols to be treated as texts. The think-aloud protocols were coded according to the following procedure: First, a set of rules was written which defined which propositional structures met the criteria of an object and which propositional structures met the criteria of an operation. The propositional structures used in this set of rules were elements from Frederiksen's (1975, 1986) theory of propositional representation for natural language. Second, the semantic structure of the protocols was analyzed to identify those propositional structures which met the criteria for an object and an operation.

Analysis of Rater Objects

The objects which were coded by the Task Independent Coding were analyzed in order to identify categories of knowledge that the raters used in making their decision. Objects were categorized according to their relation to the following: the scoring rubric (i.e., a rubric object), the content of the text (i.e., a content object), the author (i.e., an author object), the syntactic or semantic structure of the text (i.e., text object), or none of the above (i.e., other object). The total number of objects used in each of the categories by a rater was tabulated for each text. A detailed description of categories of objects is provided in Figure 4.

Analvsis of Rater Operations

The operations specified in this coding included *rater goals* (e.g., I am going to look at the rubric again), *evaluations* (i.e., an object and an attribute paired together such as [details: elaborated]; [purpose: clear]; [tone: appropriate]; [organization: cohesive]), and *relations* which are constraints on pairs of evaluations. Three types of relations were coded. The first relation-type was a conditional relation which consisted of two evaluations linked by markers in the text such as *so* or *because* (i.e., [purpose: clear] [details: elaborated] *because*). The second relation-type was an adversative conditional relation which consisted of two evaluations linked by markers in the text such as *but* (i.e., details: repetitive] [details: elaborated] *but*). The third relation-type was an OR relation which consisted of two evaluations linked by the marker *or* in the text (i.e., [purpose: Frequently] [purpose: Sometimes] *or*). Operations used by a rater to evaluate each scoring criterion were coded.

Analysis of Problem Behavior Graphs

The temporal order in which all objects and operations were selected by each rater was recorded in the form of a problem behavior graph for each of the five criteria per text analyzed in this research. A problem behavior graph is a node link structure in which the nodes represent the objects and the links represent the relations between object-attribute pairs. Problem behavior graphs provided the basis for the following analyses: 1) rater activity during the reading of a text, 2) rater activity subsequent to reading the text but prior to score assignment, and 3) rater activity subsequent to score assignment. These analyses made it possible to determine the task definitions of raters.

By comparing the problem behavior graphs of each rater for each criterion it was possible to determine the frequency with which raters constructed the same task definition for the same scoring criterion.

Reliability Check of Coding

Ten per cent of the think-aloud segments were coded independently by a second coder, a doctoral student in the Department of Educational and Counselling Psychology. Training was given in the application of the coding method prior to the reliability check. An agreement level of 93% was reached.

CHAPTER IV

Results

Rater Agreement

Rater agreement on text scores using the Vermont rubric was examined by means of correlation coefficients and percent of full agreement. As shown in Table 1 agreement in terms of correlation coefficients varied between .40 to .60, and in terms of percent agreement between 37% to 43%.

Analysis of Rater Objects

The proportions of different objects that each rater used are presented in Table 2. The pattern of object use was largely consistent across raters and sets. The raters made the most use of rubric objects, followed by content objects, followed by text objects. This pattern suggests that the raters were constructing a task definition that linked the rubric criteria with the content of the individual texts. The exception to this pattern of object use is seen for Pat with the first set of protocols: For this set, which was made up of texts from students she taught, Pat made the most use of author objects, followed by rubric and content objects.

Analysis of Problem Behavior Graphs

Analysis of rater operations. Based on analyses of problem behavior graphs it was determined whether the operation used by the rater occurred before or after the assignment of a rubric score for a text--variation across this division signals important differences in strategies used by the raters (see below). The proportions of evaluation and relation operations that each rater used before and after score assignment are presented in Table 3.

From Table 3 it can be seen that the proportion of evaluation operations was greater than that of relation operations for all three raters on both sets. Operations vary a great deal, however, in *when* they were used in relation to the actual assignment of a rubric score. For Tom the great majority of operations occurred *before* the score assignment. In contrast, for Pat most operations *followed* the score assignment. Kathy showed a mixed pattern: On the first set of protocols most of her operations occur before the score assignment, although the pattern is not as marked as Tom's; on the second set of protocols most of her operations follow the score assignment.

Rater activity while reading a text. Analyses of problem behavior graphs showed that Pat and Kathy consistently read a text without interruption and then evaluated the first criterion (i.e., purpose) in both sets. In contrast, Tom interrupted his reading of a text to evaluate its features 50% of the time in Set A. He read each text without interruption and then evaluated the first criterion 100% of the time in Set B .

Rater task definitions. Three types of task definition emerged from the analyses of the problem behavior graphs. The first task definition is considered to be a *search* process. A search process was present when the rater considered one or more alternative scores prior to score assignment. Evaluation and relation operations selected before score assignment served to facilitate the search for a solution by ruling out an alternative score (or scores). The remaining task definitions are considered to be recognition processes. A recognition process was present when the rater assigned a score without considering one or more alternative scores. Two types of recognition strategies were found. The first is a *simple recognition* process in which the rater assigned a score without first analyzing the criterion being evaluated. The second is a *complex recognition* process in which the rater analyzed the criterion being evaluated prior to score assignment. Relation and evaluation operations selected after score assignment served to justify (implicitly or explicitly) the score assigned. An example of each of these three task definition is presented below.

In the following example taken from Set A, Pat was a 'teacher as rater'. She used a simple recognition task definition when she evaluated the organization of a student's "Letter to the Reviewer" (Text 4A.3 in Appendix 111). A problem behavior graph which details the sequence of the selection of objects and operations used as part of this task definition is presented in Figure 5. (Author objects are underlined.) Pat stated,

> For organization we would give <u>Jason</u> a Frequently. It is relatively organized but <u>his</u> paragraphs are very short. <u>His</u> first and <u>his</u> last paragraph are not really highly organized paragraphs. We would not give <u>him</u> a Sometimes however because <u>he</u> does not shift in point of view and <u>he</u> does not have inconsistencies in coherence.

From the problem behavior graph it can be seen that Pat assigned a score directly without analyzing the organization of the text and without considering an alternate score. She justified score assignment indirectly by a) using an adversative conditional relation signaled by the marker <u>but</u> and b) by using evaluations not contained in the scoring rubric. She provided further justification when she used a conditional relation signaled by the marker <u>because</u> to rule out the assignment of a lower score, Sometimes.

Pat took the perspective of the author rather than the text throughout this evaluation. She selected six author objects. She assigned a score to the author rather than to the text and she evaluated that the author had inconsistencies in coherence, not that the text had inconsistencies in coherence.

In the following example taken from Set B, Tom was a 'teacher as rater'. He used a search task elaboration when he evaluated the voice and tone of the text "Vinnie" (Text 2B.2 in Appendix III). A problem behavior graph of this rating activity is presented in Figure 6. He stated,

> Voice and tone. Evidence of beginning sense of voice, some evidence of appropriate tone. Little or no evidence of voice. I think there is little or no voice evident here in this piece. It is just kind of empty. "I am writing about my best friend Vinnie". "I like the story Vinnie because it is about a homeless person who finds a home". "I shared my piece with Graham and he thought it was good". I think there is little or no voice evident here. Rarely.

From the problem behavior graph it can be seen that Tom read descriptors associated with the ratings Sometimes and Rarely and then selected an evaluation consistent with the rating Rarely. He used content objects to justify this evaluation and then repeated the evaluation prior to assigning the score Rarely.

In the following example taken from Set B, Kathy was an external rater. She used a complex recognition task definition when she evaluated the organization of the text "Hon

Yost" (Text 1B.3 in Appendix III). A problem behavior graph of this rating activity is presented in Figure 7. She stated,

> For the organization there definitely are poor transitions and I felt the shift in the point of view. He is trying to tell us that the British fell for the scam but he keeps on using the words stupid and dumb. He is telling at the end that "Hon Yost was dumb enough to do what the Americans said". "So that ends my story of the stupid Tory, Hon Yost". Sometimes under organization.

From the problem behavior graph it can be seen that Kathy selected evaluations consistent with the rating Sometimes and used content objects to provide an inferred justification for these evaluations. This rating activity also served to provide an analysis of aspects of the organization of the text as outlined by the rubric, namely transitions and shifts in point of view. All of this rating activity preceded score assignment.

Proportions of the types of task definitions used by each of the raters are presented in Table 4. Pat showed a preference for the construction of a simple recognition task definition in Set A, the set which she evaluated as 'teacher as rater'. Tom used a complex recognition task definition and a search task definition with near equal frequency in Sets A and B. Proportionately, he used each of these task definitions nearly twice as often as he used a simple recognition task definition in both sets. Kathy did not show a preference for a particular task definition in Set A yet she showed a very strong preference for the use of a simple recognition task definition in Set B. As stated earlier, unlike Pat and Tom, Kathy was an external rater in both sets.

Consistency of rater task definition per scoring criterion. The extent to which raters constructed the same task definition when evaluating the same scoring criterion is presented in Table 5. The proportion ranged from .00 to .18 in Set A and from .00 to .30 in Set B. Presented below are a series of examples which demonstrate pairs of raters using different task definitions to evaluate the same scoring criterion.

Pat was the 'teacher as rater' and Kathy was an external rater when they evaluated independently the details of a student's "Letter of Introduction" (Text 3A.1 in Appendix III). The problem-behavior graph of this rating activity is presented in Figure 8. (Author objects are underlined).

Pat stated,

> For details, we have to give Jerry a Frequently because he has lots of details. He tells us why he likes to draw. He tells us that he likes this class because we do Trivial Pursuit. And even though he didn't do all the assignments he liked the assignments. He tells us what he likes to eat and even shows us his sense of humor with "Hope to see you little people in the halls".

Kathy stated,

> Under details they certainly lack elaboration. Let's see. Are they random? See it seems to me that they are random, inappropriate or barely apparent. He's supposed to be writing a Letter of Introduction introducing himself and there is just this conglomeration that the best part is when he's talking about the drawing. How much he likes the drawing. Then we talk about Trivial Pursuit, the assignment, he doesn't like cooked peas or cooked potatoes, some

> microwave food. "You will like this class, I know I do". "Hey, see you little people in the halls. Bye". Boy, so I'm right now between Rarely and Sometimes. I just can't make a decision right now. The details certainly are random. Under Sometimes it says lack elaboration, details lack elaboration or are repetitious. Well, they are not really repetitious. When he talks about the drawing it's with some elaboration. Well, it is a couple of sentences which is more than the rest. Hmm. I don't know why I am having a little struggle with this. Well, I am going to go to Sometimes because there are some details here and maybe it is his style, having random style.

Pat used a simple recognition task definition. She assigned a score directly without analyzing the details in the text and without considering an alternate score. She used an evaluation which was not associated with the scoring rubric to justify score assignment (i.e., [details, present (elided) 'lots of']). She used content objects to illustrate this evaluation.

Kathy used a search task definition when she evaluated the details of the same text. She considered whether the details were consistent with the rating Rarely or Sometimes. She selected content objects as she worked to make a match between her response to the details of the text and the language of the scoring rubric. When she could not make this match, she assigned a score and used an evaluation which was not associated with the scoring rubric to justify score assignment (i.e., [details, 'present' (elided) 'some']).

Pat was the 'teacher as rater' and Tom was an external rater when they evaluated independently the organization of a student's "Letter about the Best Piece" (Text 2A.3 in

Appendix III). The problem behavior graph of this rating activity is presented in Figure 9. (Author objects are underlined.)

Pat stated,

> For organization, her organization would follow along a Frequently line. She is organized. She has minor lapses. She gets a little bit carried away because she is so personally involved and she wants us all to know how much it will help us to write something sad. Holly is an extremely kind, loving girl and this begins to be obvious in this piece.

Tom stated,

> Organization. The focus is pretty strong here, pretty good, I think. Starting with the general statement of the dog, his death, the special meaning to the writer, then going into how the spirit of an animal can live on in a piece of writing or a person. Fluent, cohesive. I think the second paragraph really is a nice example of fluency- explaining something for the reader. Clear focus, yes. I'm going to say logical progression of ideas, I'm going to say Extensively. It is unusual because usually the purpose, I think purpose and organization usually are very corresponding and here they are not.

Pat used a simple recognition task definition when she evaluated the organization of this text. She assigned a score directly without analyzing the organization and without considering an alternate score. She justified score assignment by selecting evaluations consistent with the rating Frequently and by using author objects.

Tom developed a complex recognition task definition when he evaluated the organization of the same text. He read multiple descriptors associated with the rating Extensively and used content objects as part of his analysis of the focus and fluency of the text prior to score assignment.

Kathy and Tom were external raters when they evaluated independently the purpose of a student's Best Piece, "Hiking to the Top" (Text 3A.2 in Appendix III). A problem behavior graph of this rating activity is presented in Figure 10.

Kathy stated,

> Okay, a tiny little adventure here. I'm glad I wasn't the teacher. Okay, so for purpose, Frequently. He establishes a purpose when he's talking about hiking Belevedere Mountain. He develops an awareness of his audience and task. Develops ideas but they may be limited in depth. But certainly, I see that he does a lot of telling and not really showing.

Tom stated,

> Okay, I don't think the purpose is clear here. Hiking to the Top is the title and the piece really isn't about that. It is, oh, this is a piece where the center of gravity really comes in these middle paragraphs. For me, it is very heavy right there. That's the heart of the piece and I don't feel, I guess I don't feel the writer has a sense of that as being what the purpose should be. Here is a place where these two paragraphs say this should be the purpose because they are so strong, so interesting. And at the beginning and the end are really, part of the process where you can get into it and get out of it but in the final drafts

> they wouldn't really be all that relative or necessary. So, does the writer attempt to establish a purpose or does the writer not establish a clear purpose. I'm going to say the writer does not establish a clear purpose here. I think he was trying to, I think the writer was thinking I am going to tell the story about what happened that day and that is it, rather than having a more focused goal. Rarely.

Kathy developed a simple recognition task definition when she evaluated the purpose of this text. She assigned a score directly without analyzing the purpose and without considering an alternate score. She used evaluations to justify score assignment.

Tom developed a search task definition when he evaluated the purpose of the same text. He selected an evaluation associated with the rating Rarely and used text objects when he analyzed the purpose of the text. He queried if the text was consistent with the rating Sometimes or Rarely. He selected an evaluation associated with the rating Rarely and content objects to illustrate this evaluation prior to score assignment.

Pat and Kathy were external raters when they evaluated independently the details of a student's Best Piece "Malachia" (Text 3B.2 in Appendix III). A problem behavior graph of this rating activity is presented in Figure 11.

Pat stated,

> Details. I'd give it a Frequently. They certainly are elaborated and they are appropriate. But they could be even better especially in the beginning. The details were better at the end. I would have liked a few more details about Malachia itself but we'd give it a Frequently.

Kathy stated,

> For the details, well, they lack a personal awareness. Repetition. It just seems that "It was a miracle that Mayor Hall had been elected". "He was a low life" but there is nothing backing that up. And also "Many said that he was a bigger loss than Gupta". It is like she is name dropping but we don't know, I don't know who these people are. I am going to say Sometimes on the details.

Pat developed a simple recognition task definition when she evaluated the details of this text. She assigned a score directly without analyzing the details and without considering an alternate score. She selected evaluations associated with the rating Frequently and used an adversative conditional relation signaled by the marker <u>but</u> to qualify her initial evaluation of the details.

Kathy developed a complex recognition task definition. She selected evaluations of the details and then used content objects to analyze the details within the text prior to score assignment.

Pat was an external rater and Tom was the 'teacher as rater' when they assessed independently the purpose of a student's "Letter about his or her Best Piece" (Text 2B.1 in Appendix III). A problem behavior graph of this rating activity is presented in Figure 12. (Author objects are underlined.)

Pat stated,

> This is a very sweet letter about the Best Piece. But the purpose, at least, I think <u>this person</u> knows the purpose. The <u>person</u> is saying 'I chose the story

> about Vinnie, I shared it with somebody else and they said it was good and I feel good'. I would say this teacher has helped this student understand that the Best Piece should be one you feel good about. Student ownership was very important here. I would give her a Sometimes. There is an attempt to establish a purpose. It certainly has not developed any ideas so I can't give her a Frequently even though I would like to because the ideas are not developed. It is not even limited in depth. It is just not enough.

Tom stated,

> Here is a Letter about the Best Piece. He's writing about his Best Piece and that is the purpose and he doesn't say much about it at all. I would say it is either rudimentary development of ideas or lacks clarity of ideas. I think it is so brief that I would say it lacks clarity of ideas. Demonstrates a minimal awareness of audience and task. I'm going to say Rarely for purpose.

Pat developed a complex recognition task definition when she evaluated the purpose of this text. She used content objects to analyze the purpose of the text prior to score assignment. Pat used an evaluation to justify score assignment and a conditional relation signaled by the marker because to rule out assignment of a higher score.

Tom developed a search task definition when he evaluated the purpose of the same text. He read descriptors associated with the ratings Sometimes and Rarely. He selected an evaluation associated with the rating Rarely (lacks clarity of ideas) based on the length of the text and then he selected a second evaluation associated with the same rating prior to score assignment.

Kathy was an external rater and Tom was the 'teacher as rater' when they assessed independently the organization of the text "Malachia" (Text 3B.2 in Appendix III). The problem behavior graph of this rating activity is presented in Figure 13.

Kathy stated,

> The organization. Here again Frequently. Organized but may have minor lapses in unity or coherence. Transition is evident. Usually has a clear focus. I got confused when her speech ended. She has three stars there to kind of show the transition, I guess, but I mean it was hard for me to follow here. It seems as if there, it is like she knows where she is going, but the audience, she is not bringing me along with her.

Tom stated,

> For organization, I think there are some transitions here that are hard to follow. Pieces that are not filled in very much. I'm not sure why, what they were expecting of this speech of Mayor Hall and why they were so upset. I could see why they were upset but I'm not sure what they were expected to do. There are other places I don't really understand. Let me see, inconsistencies in unity and/or coherence, poor transitions. I have some real problems with the coherence here in this piece. Serious errors in organization. Thought patterns difficult if not impossible to follow. Most of my other questions get cleared up as I go through the rest. I'm going to say Sometimes for organization.

Kathy developed a simple recognition task definition when she evaluated the organization of this text. She assigned a score directly without analyzing the organization and without considering an alternate score. Kathy selected evaluations to justify score assignment and she used content objects to illustrate these evaluations.

Tom developed a search task definition when he evaluated the details of the same text. He used content objects to analyze the organization and then he selected descriptors associated with the ratings Sometimes and Rarely in order to match his analysis of the organization with the language of the scoring rubric. This rating activity preceded score assignment.

CHAPTER V.

Discussion and Implications

This study identified three specific task definitions that highly experienced raters constructed when they evaluated student writing using an analytic scoring rubric. These results extend the findings of earlier research which found that proficient raters read an essay from beginning to end, without interrupting the reading to comment on the essay's content while less proficient raters seemed to go through an alternating cycle of reading and monitoring portions of the essay (Huot, 1993; Pula & Huot, 1993; Wolfe & Ranney, 1996; Wolfe, 1997). However, these earlier studies did *not* identify the task definitions of proficient raters. The identification of the different task definitions constructed by raters sheds additional light on how raters make decisions about student writing, and provides further evidence which dispels earlier beliefs that raters tacitly internalize a set of criteria which they apply directly to student writing (White, 1984). In fact, results of this research show that raters, regardless of the task definition they construct, engage in extensive problem-solving activity.

What emerged from the analyses of the think-aloud data is that a simple recognition task definition most resembled general impression scoring. When raters used a recognition task definition, they assigned a score directly. That is, they did not reread the text, analyze the features under investigation in the text, or consult the scoring rubric prior to score assignment. Additional evidence that this task definition represented general impression rating was found in the audiotapes where each pause in raters' think-alouds was recorded. Consistent with general impression scoring, when raters used a simple recognition task definition to assess the first criterion, purpose, they did not pause or engage in any other rating activity prior to score assignment. Likewise, when raters used this same task definition to assess any of the subsequent criteria, they did not pause or engage in any other scoring activity prior to score assignment. Furthermore, out of the 360 scores which were assigned by raters in this study, on only three occasions when raters constructed a simple recognition task definition (n=161) did they change their minds subsequent to score assignment.

In addition, on the basis of evidence taken from analyses of think-aloud data a distinction was made between text-based and rubric-based evaluation. A complex recognition task definition involved analysis of scoring criteria prior to score assignment but did not involve search of the rubric as part of the scoring process. As such, this task definition represented text-based evaluation because of the rater focus on analysis of specific text features. In contrast, a search task definition involved extensive search of the rubric as the rater worked to match his or her response to the text (and possibly analysis of the scoring criterion) with the scoring rubric. This task definition thus represented rubric-based evaluation because use of the rubric was central to the rater's evaluation process. Thus, the different task definitions identified in this research represent three very different foci: general impression scoring, text-based evaluation, and rubric-based evaluation.

It was interesting that there was no evidence whatsoever in the raters' protocols that they themselves were aware that they had a repertoire of different task definitions. This is an unexpected result which seems to indicate that despite their training and level of expertise, these three raters on their own developed the same three task definitions. Future research in this area will investigate if a second set of raters constructs these same task definitions or a different set.

That raters construct and apply different task definitions has important implications for the validity of writing assessment procedures. First, it appears that different task definitions affect the meaning of the scores assigned. These results indicate that it is possible for a given rater to focus on either his or her general impression of a text, the language of the rubric, or an analysis of the text relevant to the criterion being evaluated at the time. Thus, when raters take such a different focus when evaluating the same criterion, the scores they assign no longer have the same meaning. In the example above when Pat and Kathy assessed the details of a student's Letter of Introduction (p. 61), their rating activity was strikingly different. Kathy evaluated the degree of elaboration of the details and whether or not they were random. In contrast, Pat did not consider these or any other descriptors and assigned a score which she justified by using an evaluation not found in the rubric.

Second, there are two important validity issues which are usually conflated in holistic and analytic assessment. The first issue concerns whether the score is a valid assessment of the rater's response to the text. For example when Kathy analyzed the organization of the text Malachia (p. 68) she assigned a score directly and then justified score assignment. However, based on the objects and operations she used subsequent to score assignment, it appears that the rubric language associated with the rating 'Frequently' does not validly reflect her judgment of the text's organization. The second issue concerns whether the score is a valid assessment of text characteristics. That is, do the rubric guidelines adequately characterize lexical, syntactic and semantic characteristics of a text or do the guidelines offer highly-abstracted and not widely-understood concepts? This problem is illustrated in the example presented above when Kathy and Tom evaluated the purpose of the text "Hiking to the Top" (p. 64). Tom and Kathy differed in their interpretation of the criterion purpose, a difference which led Tom to conclude that "the writer does not establish a clear purpose here" (i.e., a Rarely), and Kathy to conclude that "he establishes a purpose when he's talking about hiking Belvedere Mountain" (i.e., a Frequently).

The design of this study also lent itself to the investigation of how a rater defined the rating task as an external rater and as a 'teacher as rater'. As such it was possible to examine a) the consistency with which external raters constructed the same task definition and b) the effect of knowledge of the student on rater task definition. Concerning the consistency with which external raters constructed the same task definition, as seen in Table 5, these results indicated that the proportion of the time which the external raters in Set A (Tom and Kathy) constructed the same task definition for the same task ranged from .08 to .10. The proportion of the time which the external raters in Set B (Pat and Kathy) constructed the same task definition for the same task ranged from .01 to .30.

Concerning the effect of knowledge of the student on rater task definition, these results indicated that Tom constructed and maintained the same task definitions whether he was the external rater or teacher-as-rater. Pat made a shift from showing a preference for using a simple recognition task definition when she evaluated the texts written by her students to a preference for using a complex recognition task definition when she was an external rater. Kathy who was an external rater in both sets did not maintain the same task definitions across sets, thus indicating that highly-experienced raters do not always have fixed task definitions.

As seen in Table 2, when compared to the other raters, Pat showed a preference for the selection of author objects, particularly when she was the teacher of the student writers. This preference was noted in most of the examples of her rating activity cited above but was perhaps the most salient when she evaluated the organization of a student's Letter about the Best Piece (p. 63). When compared to the other raters, Tom selected fewer author objects when he was teacher-as-rater than he did as an external rater, although this is not a statistically significant difference. This is an important finding indicating that the selection of author objects is related to individual characteristics of the rater rather than simply knowledge of the student. This is best demonstrated above when Pat and Tom evaluated the purpose of a student's "Letter about his or her Best Piece" (p. 66). Pat the external rater used more author objects than Tom the 'teacher as rater'.

With regard to the reliability of scoring, these results demonstrate that it is difficult for raters to reach agreement using traditional analytic rubrics when given traditional training. As seen in Table 5 the proportion of times which raters in this research constructed the same task definition when assessing the same text was low (e.g., ranged from .00 to .30). As seen in Table 1 per cent of rater agreement was also low (e.g., ranged from 37% to 43%). The limited proportion of times that raters constructed the same task definition when assessing the same text would indicate that raters are *not* uniformly constructing a task definition based on the first text written by a student and then constructing a simple recognition task definition for the remaining texts written by the student.

Thus, these results challenge assumptions underlying existing approaches to writing assessment by showing that raters used three different task definitions and that they rarely elaborated the same task (i.e., evaluate voice and tone) using the same task definition. While it is unclear *why* each rater elaborated the same task in a different way the majority of the time, these results nevertheless indicate the wide variability in task definitions among raters, a variability which poses a serious threat to the validity of direct writing assessment.

Implications for Writing Assessment

Perhaps the most obvious implication of these results is that raters might best counteract the variability in task definitions used here and the concomitant threat to the validity of direct writing assessment by assuming a more collaborative framework for assessment. Such a collaborative framework would necessarily involve both writing assessment *formats* and writing assessment *scoring procedures* so that both would occur in an authentic situation, constructed around social and group processes. Witte et al. (1995) discuss the concept of the *communication event* within the context of an assessment of advanced ability to communicate and one can see the potential application of this concept for the assessment of one aspect of communication, namely writing:

> We see the communication event as a series of parallel but integrated activities that unfold across a protracted period of time and that involve not individuals working in isolation but a group working toward some common goal, which the group itself both identifies and defines (p. 43).

They believe it is possible to construct, validate, and apply scales of various types across diverse sources of data on communicative processes. For example, scales could be designed for rendering judgments of the appropriateness or effectiveness of students' instrumental and epistemic uses of written language, for assessing how effectively the students use language activities focused on planning or problem-solving, and for evaluating participants contributions to small and large group meetings (p.53-54). They recommend that teacher facilitators could serve as important sources of descriptive and evaluative information on the performances of the students. Students themselves would be another important source of evaluative information. Thus, this assessment format outlined by Witte et al. (1995) is highly compatible with the interpretative approach to writing assessment as advocated by Delandshere et al. (1994), Moss et al. (1992), and Moss (1996).

As these results have shown, raters who followed a psychometric approach to assessment appear to have lacked a consistent interpretation of the assessment task (both within and across raters). In contrast, it appears that an interpretative approach to assessment which builds on collaborative judgments and written summaries could be a step toward the development of a shared interpretation of the assessment task. That is, collaborative activity could promote shared interpretation of the evidence being evaluated and the language used to evaluate the evidence (Moss, 1996; Moss et al., 1992).

One possible implication of this research is to incorporate the three task definitions identified here into an interpretative approach to writing assessment. Such an approach would consist of the following steps: a) pairs of raters generate a hypothesis about the quality of the criterion by stating their general impression of the criterion (i.e., simple recognition task definition), b) the pair of raters then collaborate to write their analysis of the criterion being evaluated in the form of an interpretative summary (i.e., complex recognition task definition, c) the pair of raters then match their analysis of the criterion being evaluated with the language of the scoring rubric and justify this "match" (i.e. search task elaboration) and d) the pair of raters then compare the rating they generated at step one (i.e., general impression rating) with the rating they generated at step three (i.e., search task definition). If the rating is the same, then the pair of raters assign that score. If the rating is not the same, then the raters begin the process a second time and the rating generated at step three becomes the working hypothesis. Thus, the pair of raters generate an initial interpretation, then collaborate on an interpretative summary, then match this summary with rubric guidelines and justify this "match". They either accept the initial interpretation, or challenge it and revise it.

The same text or collection of texts is evaluated by a second pair of raters. Where there is rater disagreement, investigation of interpretative summaries and justifications of "matches" can provide the basis for a rater to mediate between the two pairs of raters.

This approach is similar to other interpretative approaches in that it assumes the centrality of written analysis of scoring criteria (in the form of interpretative summaries) and the collaborative work of raters. This approach differs from other interpretative approaches in the following ways: a) the incorporation of the three task definitions revealed in this research, b) a pair of raters who serve as their own control, and c) the process of external replication.

Raters' general impressions are used in this approach because it is has been shown consistently in past research (Condon & Hamp-Lyons, 1994; Sommers et al., 1993) and in this present research that raters have a tendency toward general impression rating which they act upon inconsistently (i.e., at chance level in this research). Thus, by formalizing the use of a general impression rating, this method enables raters to incorporate an initial hypothesis into their decision making process as an alternative to general impression rating as currently practiced. In addition, the initial hypothesis plays an integral part in enabling raters to serve as their own control by comparing their initial hypothesis with their final decision about score assignment. Concerning external replication, in the method proposed here the pair of raters compares the rating assigned to that assigned by a second pair of raters.

Implications for Training Raters

Results of this research which have shown that there is marked variability in the use of task definitions by raters have important implications for the training of raters. These results indicate that experienced raters use a variety of different task definitions, only one of these definitions being a direct matching process. Thus, the implication for training would follow the implications for writing assessment. First, it appears that raters would need to receive training in *how* to analyze evidence specific to scoring criteria, as well as training in how to write interpretative summaries based on these analyses. Second, it appears that raters would need to be trained how to match the evidence contained in the interpretative summaries with the scoring standards of the assessment method. In light of the validity concerns expressed about existing scoring criteria (Gere, 1980; Greenberg, 1992; Wiggins, 1994) it is likely that raters will need training in how to evaluate different sets of scoring criteria such as that proposed by the work of Gearhart and Wolf (1994) and Wiggins (1994). Recent research in the area of persuasive writing (Crammond, 1996) and narrative writing (Senecal, 1998) can help to guide the construction of rubrics which include genre-specific criteria. Furthermore, it is likely that future scoring rubrics following the work of Witte et al. (1994, 1995) will place more emphasis on social and communicative aspects of writing, particularly the ability to collaborate with others when writing in a given situation. Thus, training in how to interpret and apply new criteria as well as how to write an interpretative summary are important implications of this research for training raters.

Implications for Writing Instruction

Curriculum and assessment exist in a reciprocal relationship, with each influencing the other (Murphy, 1994); however, it is well-known that assessment can *drive* instruction (see Moss, 1994a; Crooks, 1988). Given that the results reported here lend support to an interpretative approach to assessment, and given the relationship between curriculum and assessment, then these results also lend support to a collaborative approach to writing instruction. Dyson and Freedman (1990) discussed methods such as peer response groups, peer writing groups, and community rather than school-based writing which enable teachers to provide students with a variety of kinds of social interaction around writing. This social interaction around writing includes interactions between students and teacher as well. They advocate a support system for writing development which enables teachers to be sensitive to their students' current skills and understanding and to provide collaborative support to help them develop further. Freedman (1987) uses the term collaborative problem-solving to try to capture the dynamic role of interaction in the process of teaching and learning writing.

Producing and using texts are always in some sense collaborative acts (Witte et al., 1994) and extensive research has documented the positive effects of student collaboration in writing instruction (see O'Donnell, Dansereau, Rocklin, Lambiotte, Hythecker, & Larson, 1985), yet as noted by Wiggins (1994) there is very little, if any, attention given to *evaluating* one's ability to collaborate during writing. Thus, if validity standards such as 'transparency' in Frederiksen and Collins's Principles of Systemically Valid Testing (1989) are to be used, then the scoring criteria which is to be made transparent to student writers should include criteria related to collaboration during writing. In this way, if assessment does guide instruction, and if the assessment is built around the construct as recommended by Messick (1994), then the reciprocal relationship between curriculum and assessment can hopefully be a healthier, more productive relationship than exists currently.

Limitations of the Study

Two possible limitations of this research are related to methodological considerations. First, the use of a case study approach with three raters as the individual cases may appear to seriously limit the generalizability of these results. However, the decision to select three raters who were highly trained in the use of an existing scoring rubric was based on the ill-structured nature of the writing assessment task. That is, given the range of solutions possible in the solving of ill-structured problems, a case study methodology was the most appropriate way to begin to study questions of within-rater and between-rater consistency in solving a number of problems (see Voss & Post, 1988). In addition, by having a limited number of raters it was possible to track each raters' *sequence* of rating activity for each text evaluated. Previous think-aloud studies have been unable to provide this particular insight into the rating process. Second, when a think-aloud methodology is used there are usually concerns raised about the validity and completeness of concurrent verbal reports. However, Ericsson and Simon (1993) state that the information that is heeded during the performance of a task is the information that is reportable; and the information that is reported is information that is heeded (p. 167). The extent of the clausal units generated by the raters (i.e., 10,000) and the rare occurrence of pauses in the audiotapes support that the verbal data collected in this study were a valid representation of how the raters in this study used an analytic scoring rubric to evaluate writing.

References

Allen, M.S. (1995). Valuing differences: Portnet's first year. Assessing Writing, 2, (1), 67-89.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). Standards for educational and psychological testing. Washington, DC: Authors.

Anderson, J. (1990). Cognitive psychology and its implications (3rd ed.). New York: Freeman.

Baker, E.L., & Linn, R. (1992). Writing portfolios: Potential for large scale assessment. Project 2.4: Design theory and psychometrics for complex performance assessment. Design and analysis of portfolio and performance measures (Tech. Rep. No. 143). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Bracewell, R.J. (1987, April). Semantic and textual constraints students use in revising their writing. Paper presented at the American Educational Research Association, Washington, D.C..

Bracewell, R.J., & Breuleux, A. (1990, April). Problem solving models of strategy implementation in expert and pre-expert writing. Paper presented at the meeting of the American Educational Research Association, Boston, MA.

Bracewell, R. J., & Breuleux, A. (1994). Substance and romance in the analysis of think-aloud protocols. In P. Smagorinsky (Ed.), Speaking about writing: Reflections on research methodology (pp. 55-88). Newbury Park, CA: Sage Publishing.

Bracewell, R.J., & Witte, S.P., (1997, March). The implications of activity, practice, and semiotic theory for cognitive constructs of writing. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Breland, H.M., & Gaynor, J.L. (1979). A comparison of direct and indirect assessments of writing skill. Journal of Educational Measurement, 16, 119-128.

Breuleux, A. (1991). The analysis of writers' think-aloud protocols: Developing a principled coding scheme for ill-structured tasks. In G. Denhière, J.P. Rossi (Éds.), Texts and text processing (pp. 333-362). Amsterdam: North Holland.

Calfee, R. (1994a). Implications of cognitive psychology for authentic assessment and instruction (Tech. Rep. No. 69). Berkeley, CA: University of California at Berkeley, National Center for the Study of Writing.

Calfee, R. (1994b). Ahead to the past: Assessing student achievement in writing (Occasional paper No. 39). Berkeley, CA: University of California at Berkeley, National Center for the Study of Writing.

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

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

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

Cherry, R.D. & Meyer, P.R. (1993). Reliability issues in holistic assessment. In M. Williamson & B. Huot (Eds.), Validating holistic scoring for writing assessment: Theoretical and empirical foundations (pp. 109-141). Cresskill, NJ: Hampton Press.

Condon, W. & Hamp-Lyons, Liz (1994). Maintaining a portfolio-based writing assessment: Research that informs program development. In L. Black, D. Daiker, J. Sommers, & G. Stygall (Eds.), New directions in portfolio assessment: Reflective practice, critical theory, and large-scale scoring (pp. 277-285). Portsmouth, NH: Heinemann, Boynton/ Cook.

Crammond, J.G. (1996). An analysis of argument structure in expert and student persuasive writing. Unpublished doctoral dissertation. McGill University.

Crooks, T.J. (1988). The impact of classroom evaluation practices on students. Review of Educational Research, 85, 438-481.

Delandshere, G. & Petrosky, A. (1994). Capturing teachers' knowledge: performance assessment a) and post-structuralist epistemology, b) from a post-structuralist perspective, c) and post-structuralism, d) none of the above. Educational Researcher, 23 (5), 11-18.

DeRemer, M., & Bracewell, R.J. (1991, April). The hidden task in the subjective rating of student writing. Paper presented at the annual meeting of the American Educational Research Association, Chicago, Il.

DeRemer, M., & Bracewell, R.J. (1995, April). Writing portfolio assessment: Can teachers assess their own students writing reliably? Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Dyson, A. H., & Freedman, S.W. (1990). On teaching writing: A review of the literature (Occasional Paper No. 20). Berkeley, CA: Center for the Study of Writing.

Elbow, P. & Yancey, K.B. (1994). On the nature of holistic scoring: An inquiry composed on Email. Assessing Writing, 1 (1), 91-108.

Ericsson, K.A. & Simon, H.A. (1993). Protocol analysis: Verbal reports as data (Rev. ed.). Cambridge, MA: MIT Press.

Flower, L.S., Schriver, K., Carey, L., Haas, C., & Hayes, J.R. (1989). Planning in writing: The cognition of a constructive process. (Tech. Rep. No. 34). Berkeley, CA: Center for the Study of Writing.

Frederiksen, C.H. (1975). Representing logical and semantic structure of knowledge acquired from discourse. Cognitive Psychology, 7, 271-348.

Frederiksen, C.H. (1986). Cognitive models and discourse analysis. In C.R. Cooper and S. Greenbaum (Eds.), Written communication annual: An international survey of research and theory: Vol. 1: Linguistic approaches (pp. 227-267). Beverly Hills, CA: Sage.

Frederiksen, C., Bracewell, R.J., Breuleux, A., & Renaud, A. (1990). The cognitive representation and processing of discourse: Function and dysfunction. In Y. Johanette & H. Brownell (Eds.), Discourse ability and brain damage: Theoretical and empirical perspectives (pp. 69-110). New York: Springler Verlag.

Frederiksen, J.R., & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18 (9), 27-32.

Gearhart, M., Herman, J.L., Baker, E.L., & Whittaker, A.K. (1992). Writing portfolios at the elementary level: a study of methods for writing assessment (Tech. Rep. No. 337). Los Angeles, CA: University of California, Center for the Study of Evaluation.

Gearhart, M. & Wolfe, S.A. (1994). Engaging teachers in assessment of their students' narrative writing: The role of subject matter knowledge. Assessing Writing, 1 (1), 67-90.

Gere, A. (1980). Written composition: Toward a theory of evaluation. College English, 42, 44-58.

Godshalk, F., Swineford, F., & Coffman, W. (1966). The measurement of writing ability (Research Monograph No. 6). New York: College Entrance Examination Board.

Gouvernement du Québec, Ministère de l'Èducation (1990). Student writing and its correction (Publication No. 1990-9091-8035). Montréal, QC: Author.

Greenberg, K.L. (1992).Validity and reliability issues in the direct assessment of writing. WPA: Writing Program Administration, 16 (1-2), 7-22.

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

Hayes, J.R. (1989). The complete problem solver (2nd ed.). Hillsdale, NJ: Erlbaum.

Hayes, J.R., & Flower, L.S. (1980). Identifying the organization of writing processes. In G.L. Gregg and E. Steinberg (Eds.), Cognitive processes in writing: An interdisciplinary approach (pp. 31-50). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

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

Huot, B. (1990b). Reliability, validity, and holistic scoring: What we know and what we need to know. College Composition and Communication, 41 (2), 201-211.

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

Koretz, D.M., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1992). The reliability of scores from the 1992 Vermont portfolio assessment program (Interim Report). Santa Monica, CA: Rand Institute on Education and Training; and Los Angeles: National Center for Research in Evaluation, Standards, and Student Testing.

LeMahieu, P. G., Gitomer, D.H., & Eresh, J.T. (1995). Portfolios in large-scale assessment: Difficult but not impossible. Educational Measurement, 14 (3), 11-28.

Linn, R. (1994, December). Performance assessment: Policy, promises, and technical measurement standards. Educational Researcher, 23 (9), 4-14.

Linn, R., Baker, E.L., & Dunbar, S.B.(1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20 (8), 5-21.

Lloyd-Jones, R. (1977). Primary trait scoring. In C.R. Cooper & L. Odell (Eds.), Evaluating writing: Describing, measuring, and judging (pp. 3-32). Urbana, IL: NCTE.

Mehrans, W. (1992). Using performance assessments for accountability purposes. Educational Measurement: Issues and Practice, 11 (1), 3-20.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher, 18 (2), 5-11.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23 (2), 13-24.

Meyer, C.J. (1992). What's the difference between authentic and performance assessment? Educational Leadership, 49 (8), 39-40.

Miles, M. & Huberman, A.M. (1984). Qualitative data analysis: A sourcebook of new methods. Beverly Hills: CA, Sage Publications.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research, 62 (3), 229-258.

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

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

Moss, P.A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretative research traditions. Educational Researcher, 25 (1), 20-28.

Moss, P.Á., Beck, J.S., Ebbs, C., Matson, B., Muchmore, J., Steele, D., Taylor, C., & Herter, R. (1992). Portfolios, accountability, and an interpretative approach to validity. Educational measurement: Issues and practice, 11 (3), 12-21.

Murphy, S. (1994). Portfolios and curriculum reform: Patterns and practice. Assessing Writing, 1 (2), 175-206.

Myers, M. & Pearson, P.D. (1966). Performance assessment and the literacy unit of the New Standards Project. Assessing Writing, 3 (1), 5-29.

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.

O'Donnell, A.M., Dansereau, D.F., Rocklin, T., Lambiotte, J.C., Hythecker, V.I., & Larson, C.O. (1992). Cooperative writing: Direct effects and transfer. In J.R. Hayes, R.E. Young, M.L. Matchett, M. McCaffrey, C. Cochran, & T. Hajduk (Eds.), Reading empirical research studies: The rhetoric of research (pp. 374-381). Hillsdale, NJ: Lawrence Erlbaum.

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

Purves, A.C. (1992). Reflections on research and assessment in written composition. Research in the Teaching of English, 26 (1), 108-122.

Resnick, L.B., & Resnick, D. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford & M.C. O'Connors (Eds.), Cognitive approaches to assessment (pp. 37-75). Boston: Kluwer Academic Publishers.

Resnick, L. B. & Resnick, D. (1993). Report on performance standards in mathematics and English: Results from the New Standards Project, Big Sky Scoring

Conference. Project 2.3: Complex performance assessments: Expanding the scope and approaches to assessment. Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing and Pittsburgh, PA: Learning Research and Development Center.

Senecal, L. (1998). The idea structure of students' written stories in grades three to five. Unpublished doctoral dissertation. McGill University

Simon, H. (1978). Information-processing theory of human problem solving. In W. K. Estes (Ed.), Handbook of learning and cognitive processes: Vol. V. Human information processing (pp. 271-295). Hillsdale, New Jersey: Erlbaum.

Sommers, J., Black, L., Daiker, D., & Stygall, G. (1993). The challenge of rating portfolios: What WPAs can expect. WPA: Writing Program Administrator, 17 (1-2), 7-29.

Spandel, V. & Stiggens, R.J. (1980). Direct measures of writing skill: Issues and applications. Portland, OR: Northwest Regional Educational Development Laboratory.

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. 111-125). Norwood, NJ: Ablex.

Vermont Department of Education. (1991). "This is my best"-Vermont's writing assessment program. Montpelier, Vt: Author.

Voss, J. F. & Post, T.A. (1988). On the solving of ill-structured problems. In M. Chi. R. Glaser, and M.J. Farr (Eds.), The nature of expertise (pp. 261-285). NJ: Lawrence Erlbaum Associates.

White, E.M. (1984). Holisticism. College Composition and Communication, 35, 400-409.

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

Witte, S.P. (1988, March). Writing prompts and composing. Paper presented at the Conference on Writing Assessment, Minneapolis, MI.

Witte, S.P. (1993). No guru, no method, no teacher: The communication domain and NACSL (Rev. ed.). Paper commissioned by the National Center for Educational Statistics for the November, 1992, Study Design Workshop on the National Assessment of College Student Learning.

Witte, S.P., & Flach, J. (1994). Notes toward an assessment of advanced ability to communicate. Assessing Writing, 1, (2), 207-246.

Witte, S.P., Flach, J., Greenwood, C., & Wilson, K.E. (1995). More notes toward an assessment of advanced ability to communicate. Assessing Writing, 2, (1), 21-65.

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

Wolfe, E.W., & Feltovich, B. (1994, April). Learning how to rate essays: A study of scorer cognition. Paper presented at the American Educational Research Association, New Orleans, LA.

Wolfe, E.W., & Ranney, M. (1996). Expertise in essay scoring. In D.C. Edelson and E.A. Domeshek (Eds.), Proceedings of ICLS 96 (pp. 545-550). Charlottesville, VA: Association for the Advancement of Computing in Education.

Table 1

Rater Agreement: N=24

Raters	Pearson r	Percent Agreement
Pat and Tom	.40	43
Pat and Kathy	.60	37
Tom and Kathy	.49	42

Table 2

Proportion of Objects Used by Raters in Set A and Set B

	Pat	Tom	Kathy
		Set A	
Rubric	0.29	0.40	0.47
Content	0.17	0.27	0.19
Author	0.35	0.09	0.14
Text	0.12	0.17	0.16
Other	0.06	0.05	0.04
(# of objects)	(n=877)	(n=1,156)	(n=834)
		Set B	
Rubric	0.37	0.52	0.46
Content	0.21	0.23	0.20
Author	0.16	0.06	0.10
Text	0.20	0.15	0.18
Other	0.06	0.03	0.06
(# of objects)	(n=779)	(n=578)	(n=564)

Table 3

Proportion of Operations Used Before and After Score Assignment in Set A and Set B

	Pat	Tom	Kathy
	Set A		
Evaluation			
Before Score Assignment	0.32	0.57	0.43
After Score Assignment	0.44	0.09	0.37
Relation			
Before Score Assignment	0.05	0.29	0.13
After Score Assignment	0.19	0.05	0.07
(# of operations)	(n=135)	(n=199)	(n=169)
	Set B		
Evaluation			
Before Score Assignment	0.25	0.70	0.37
After Score Assignment	0.42	0.13	0.51
Relation			
Before Score Assignment	0.09	0.15	0.02
After Score Assignment	0.23	0.02	0.09
(# of operations)	(n=162)	(n=109)	(n=135)

Table 4

Proportion of Raters' Task Definitions in Set A and Set B (counts of Task Definitions given in parenthesis)

	Search	Recognition (Simple)	Recognition (Complex)
		Set A	
Pat	0.00	0.65	0.35
	(0)	(39)	(21)
Tom	0.42	0.20	0.38
	(25)	(12)	(23)
Kathy	0.25	0.43	0.32
	(15)	(26)	(19)
		Set B	
Pat	0.05	0.40	0.55
	(3)	(24)	(33)
Tom	0.37	0.23	0.40
	(22)	(14)	(24)
Kathy	0.08	0.77	0.15
	(5)	(46)	(9)

Table 5

Proportion which Raters Constructed the Same Task Definition for the Same Task in Set A and Set B

Raters	Search	Simple Recognition	Complex Recognition
		Set A	
Pat and Tom	.00	.15	.10
Pat and Kathy	.00	.18	.12
Tom and Kathy	.10	.08	.08
		Set B	
Pat and Tom	.00	.10	.26
Pat and Kathy	.01	.30	.10
Tom and Kathy	.00	.30	.10

Figure 1. Relations between rater and teaching status.

	Rater		
Portfolio Set A	Pat (Teacher)	Tom (External)	Kathy (External)
Portfolio Set B	Pat (External)	Tom (Teacher)	Kathy (External)

Figure 2. The Vermont Writing Assessment - Analytic Assessment Guide (1991)

VERMONT WRITING ASSESSMENT - ANALYTIC ASSESSMENT GUIDE

	PURPOSE	ORGANIZATION	DETAILS	VOICE / TONE	USAGE, MECHANICS, GRAMMAR
ASSESSING, CONSIDER...	the degree to which the writer's response Establishes and maintain a clear purpose Demonstrates an awareness of audience and task Exhibits clarity of ideas	the degree to which the writer's response illustrates Unity Coherence	the degree to which the details are appropriate for the writer's purpose and support the main point(s) of the writer's response	the degree to which the writer's response reflects personal investment and expression	the degree to which the writer's response exhibits correct Usage (e.g tense formulation agreement, word choice) Mechanics - spelling, capitalization, punctuation Grammar Sentences as appropriate to the piece and grade level
EXTENSIVELY	Establishes and maintain a clear purpose Demonstrate a clear understanding of audience and task Exhibits ideas that are developed in depth	Organized from beginning to end Logical progression of ideas Clear focus Fluent, cohesive	Details are effective vivid, explicit, and/or pertinent	Distinctive voice evident Tone enhances personal expression	Few if any, errors are evident relative to length and complexity
FREQUENTLY	Establishes a purpose Demonstrate an awareness of audience and task Develops ideas, but they may be limited in depth	Organized but may have minor lapses in unity or coherence Transitions, evident Usually has clear focus	Details are elaborate and appropriate	Evidence of voice Tone appropriate for writer's purpose	Some errors are present
SOMETIMES	Attempts to establish a purpose Demonstrates some awareness of audience and task Exhibits rudimentary development of ideas	Inconsistencies in unity and/or coherence Poor transitions Shift in point of view	Details lack elaboration or are repetitious	Evidence of beginning sense of voice Some evidence of appropriate tone	Multiple errors and/or patterns of errors are evident
RARELY	Does not establish a clear purpose Demonstrates minimal awareness of audience and task Lacks clarity of ideas	Serious errors in organization Thought patterns difficult, if not impossible, to follow Lacks introduction and/or conclusion Skeletal organization with brevity	Details are random, inappropriate, or barely apparent	Little or no voice evident Tone absent or inappropriate for writer's purpose	Errors are frequent and severe
		Is illegible: i e., includes so many undecipherable words that no sense can be made of the response			

Figure 3. Criteria for analysis of rater objects and operations.

1. Task Analysis:

   •Explicit task instructions to Rater

   •Inferred by Experimenters (e.g., select text object)

2. Nonverbal Rater activity (e.g., reading text, reading the scoring rubric)

3. Think aloud verbalizations (analyses based on Frederiksen (1975, 1986) propositional structures):

   •Operations

   Evaluations (defined by psychological judgment by rater):

   a) Simple (ATTribute relations between objects)

   b) Comparative (ORDer relation on Attribute or DEGree between objects)

   Goal Setting (after Breuleux, 1991):

   Rater as AGENT, action which is volitional, future, and / or modalized, or queried argument

   Dependency and Logical relations:

   CONDitional, Adversative CONDitional, EXCLusive OR, EQUIValence,

   •Objects

   Rubric

   Content

   Author

   Text

   Other

Figure 4. Guideline for identification of rater objects.

•Rubric objects (R) include purpose, organization, details, voice and tone, and GUM.

•Author object (A) includes the following: the author's name, he/ she, his/her, the writer, and this student, but not reference to the author when this reference is a paraphrase of the text.

•Content objects (C) include the following:

1. Objects which result from direct reading of the student's text.
2. A paraphrase by the rater of the text. "He (A) talks about drawing (C)".
3. Statements about the writer's meaning or intention.

•Text Objects (T) include the following:

1. Direct reference to the portfolio. "We are on the third portfolio" (T)
2. Titles of texts. "This is the Letter of Introduction." (T)
3. Type of text. "He is responding in this personal narrative" (T)
4. Concrete properties of the text such as the title: "The title (T) is very good.
5. Location within the text: "He shifts around in the last paragraph "(T)
6. Direct reference to the text: "The errors are so severe that I can't read it". (T)
7. Indirect reference to the text: "I don't get much sense of purpose here". (T)
8. Reference to features in the surface structure of the text such as syntactic organization (topicalization) or vocabulary.

•Other (O) Other objects include those objects which are not rubric, portfolio, author, content and text objects.

Figure 5. Rater evaluation of the organization of "Letter of Introduction".

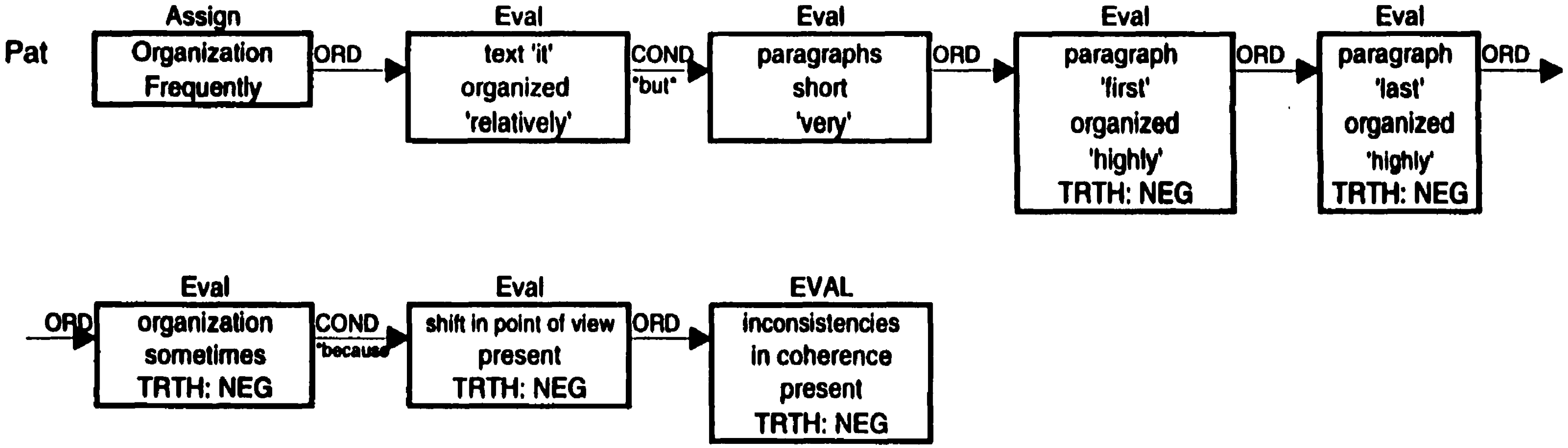


Figure 6. Rater evaluation of the voice and tone of "Vinnie".

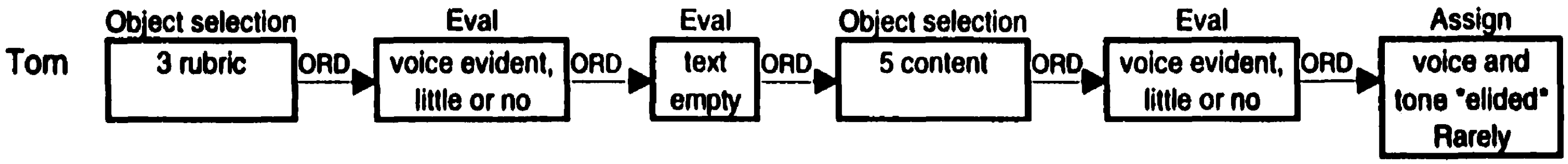


Figure 7. Rater evaluation of the organization of "Hon Yost".

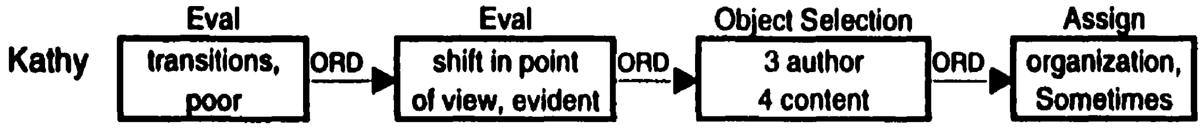


Figure 8. Raters' evaluations of the details of "Letter of Introduction".

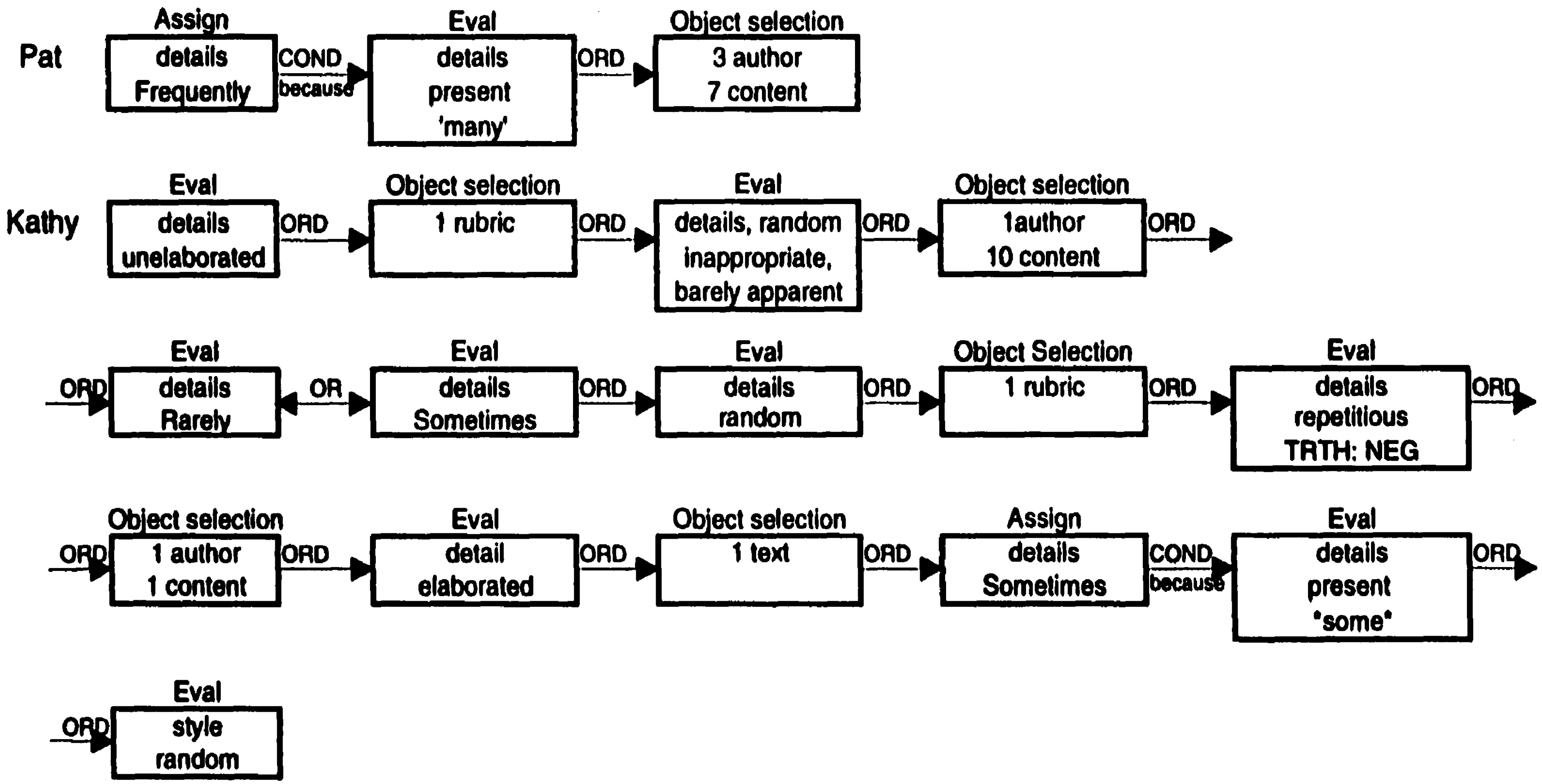


Figure 9. Raters' evaluations of the organization of "Letter about the Best Piece".

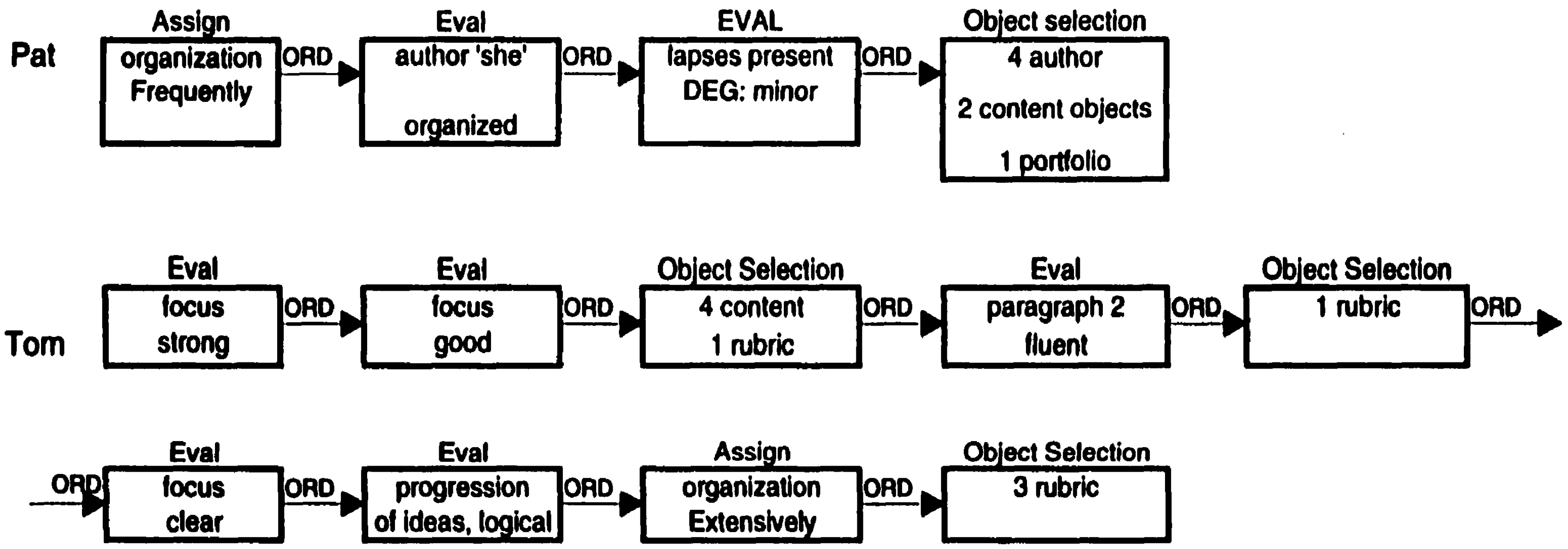


Figure10. Raters' evaluations of the purpose of "Hiking to the top".

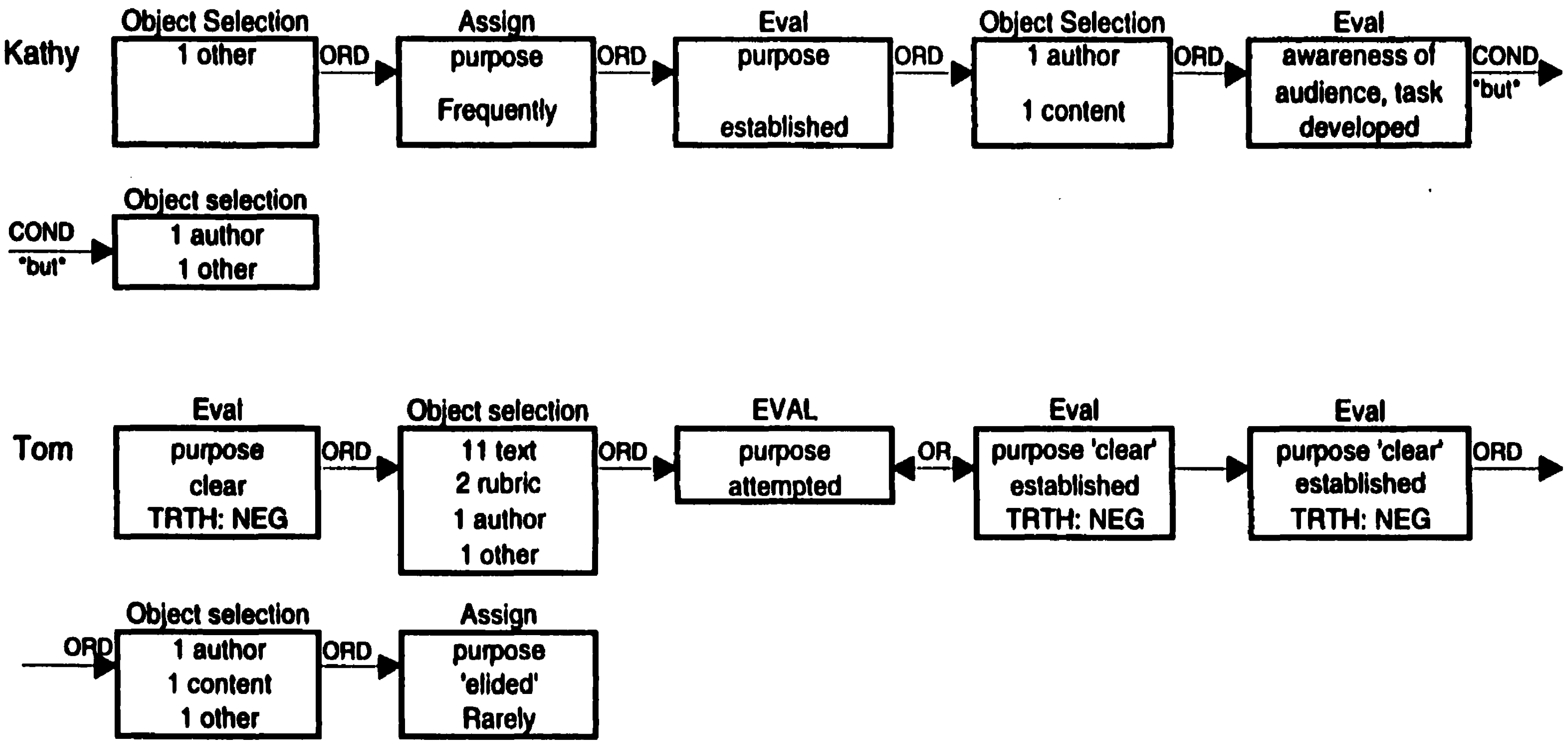


Figure11. Raters' evaluations of the details of "Malachia".

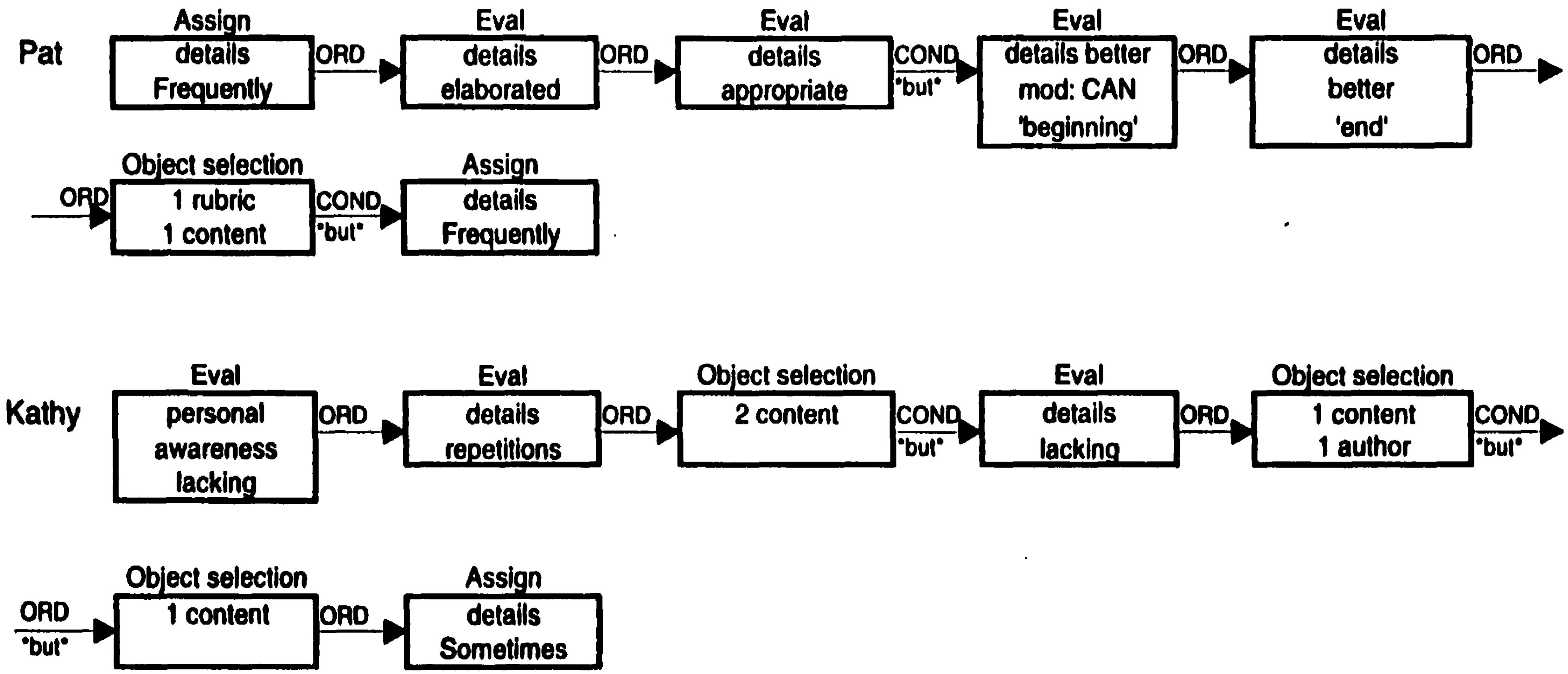


Figure12. Raters' evaluations of the purpose of "Letter about the Best Piece".

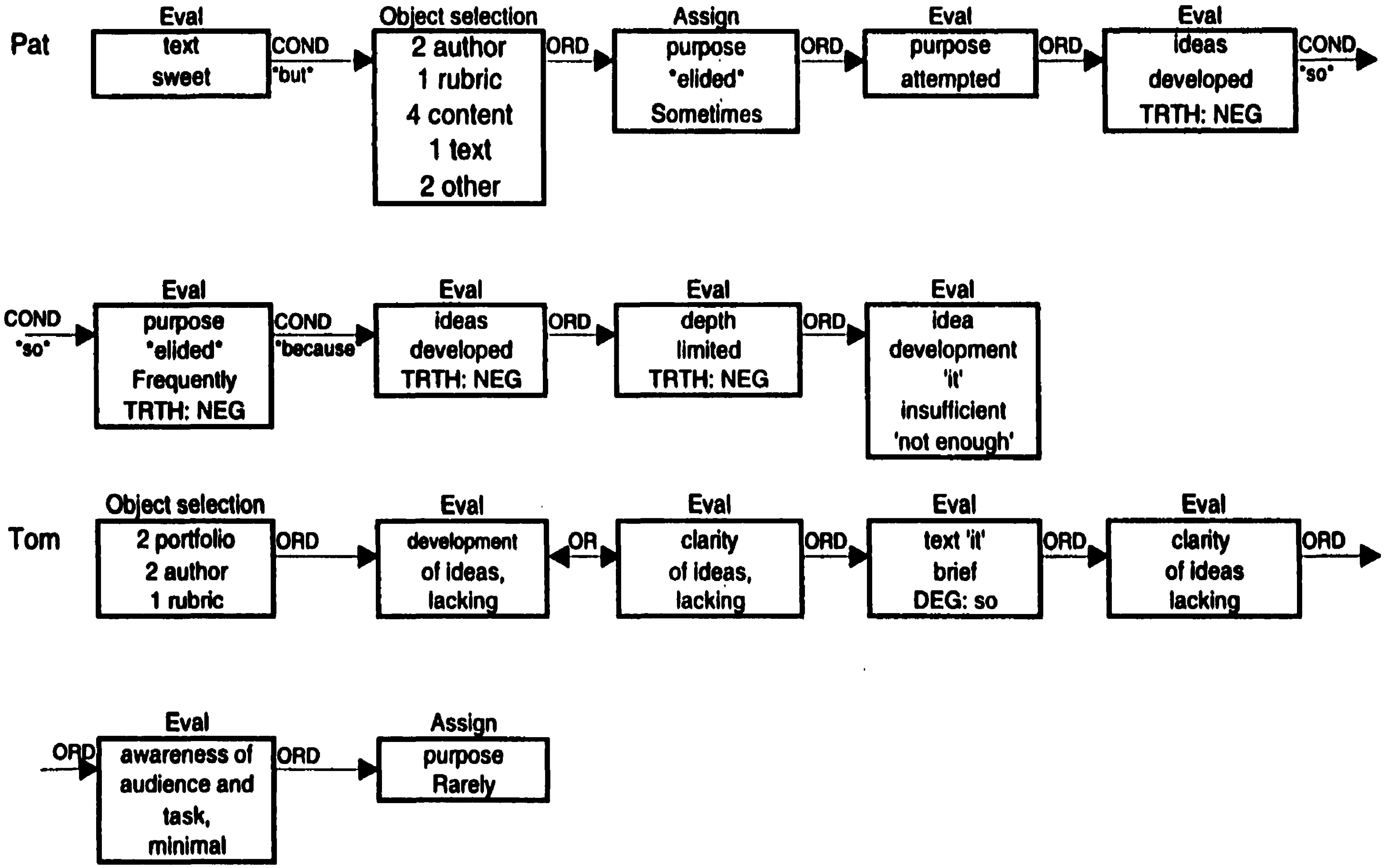
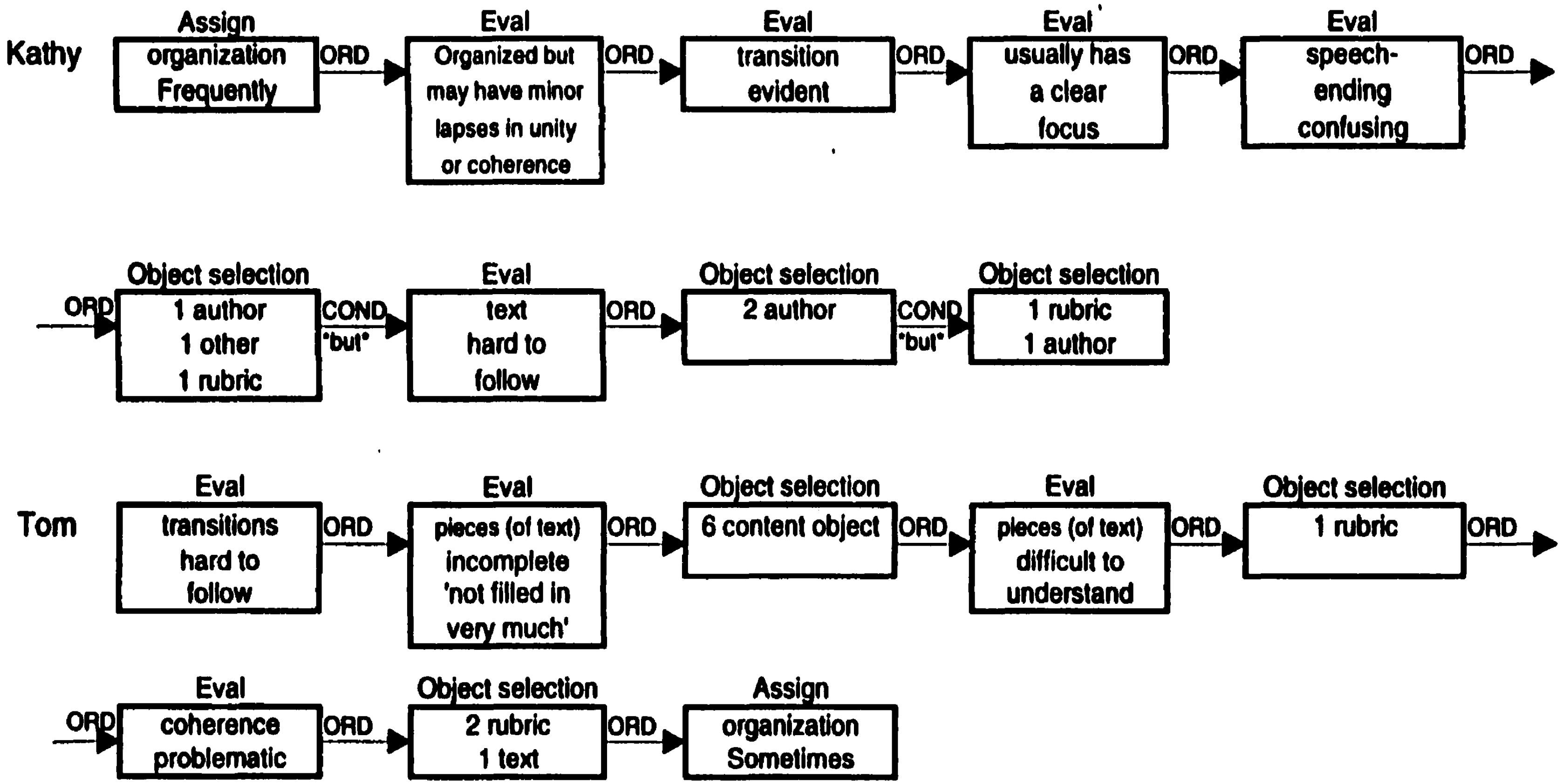


Figure13. Raters' evaluations of the organization of "Malachia".



Appendices

Appendix I

Research Ethics Committee of the Faculty of Education, McGill University

Certificate of Ethical Acceptability for Research Involving Human Subjects

Appendix II

Instructions to Raters

You will be given eight writing portfolios produced by students in grade eight. The task is to assess the first three texts of each portfolio using the Vermont Writing Assessment- Analytic Assessment Guide. Please read the texts within the portfolio in the order in which they are presented. We would like you to think aloud throughout the rating process, that is, verbalize all comments and impressions you might have as you read a text, subsequent to reading a text, and while deciding the rating to assign to a scoring criteria.

Appendix III

Student Texts

"Letter to the Reviewer" (Text 2A.3)

My favorite piece has got to be Max and I. Max was my favorite dog I ever had. He was beaughtiful. Max died a long, very painful death.

I decided to write about Max because of what he meant to me. Max was my best friend, and I wanted his spirit to live on. It makes me very happy to know that Max was loved so much and cared for.

In your life time you might have loved someone and they passed away. You might not want to let them go because of all the happy memories you had together. You want those memories you had together. It may take a long time to write this piece, but sooner or later, if you have writen your piece right the voice in it will make the loved ones spirt so alive and happy you will feel their breath against your skin. Their skin will live again.

"Letter of Introduction" (Text 3A.1)

Hello Reader,

My name is Chuck and I like to draw a lot. Drawing to me is fun. Because I can draw a lot of things and you can make any thing you want by drawing it. Ill tell you what I like about this class. By the way there not in order I liked trivial pursuit and the assignments even I didn't do all the homework. I like to eat every thing except peas and coaked patatoes and some microwave foods. You will like this class I know I do.

Hay see you littel people in the halls

By!

"Hiking to the Top" (Text 3A.2)

When our class hiked Belvidere Mountain it took so long to get to the top it felt like forever. When we got to the top there was a tower. We walked up the tower but you couldn't see anything except the fog. So we climbed down and ate lunch. After lunch we headed down the mountain.

A whole bunch of us ran down. Everyone that ran slipped, fell, and slid down the mountain. I slid off the trail. I grabbed a tree but it broke. It didn't stop me. I kept on going and ran right over a ten foot cliff. I landed on my knees. The ground was moist and the snow made it soft so I didn't get hurt.

Then I started running through the woods. I found a brook and ran down the brook until I found the trail and everyone else.

Then we boarded the bus. Everyone was soaked. We stopped at a store on the way back and brought snacks. I didn't buy anything because they didn't gave the candy bar I wanted (a Snickers). Finally we arrived at school. I had a good time.

"Letter to the Reviewer" (Text 4A.3)

Dear Reader,

My best piece is "Why Grampa?" because it has voice and detail and shows what happens when you love someone.

The interesting point is this poem was not true. I fooled everyone in my class and even the teacher. Then I read it to the principal when he came to visited the class. They all tried so hard to express sympathy. I had to tell them the truth. So I chose this piece because I thought it must have been pretty good writing. if everyone thought it was true.

I hope you had fun reading it because I had fun writing it.

Sincerely,

"Hon Yost" (Text 1B.3)

Hon Yost Shuyler, a half insane Tory, helped the Americans in a trick to make the British retreat, and caused the British Generals to conflict and end up killing each other.

Since the Americans knew that Hon Yost was a little on the dumb side, they decided to make a deal with him.

Since the Americans captured Hon Yost's brother, they made a deal with Hon to free him. They told him that if he went to the British fort and told them that the Americans had 5,000 men and were coming to defeat them, they would let Hon Yost's brother go free. So the Americans took Hon Yost's overcoat, and filled it full of bullet holes to make it look like he had been shot at, and hoped this would be more convinsing. So he set to the British fort, and when he got there he told them the made up story. The two British Generals began to conflict over what Hon Yost had told them. They fought and fought and ended up killing each other in the process. In the mean time the British were retreating as far away as possible.

The American's scam worked, and they only had three hundred men not five thousand. Hon Yost got his way also, because he cared for his brother and he was stupid enough to do what the Americans said. So that ends my story of the stupid Tory, Hon Yost.

"Letter to the Reviewer" (Text 2B.1)

Dear Portfolio Reviewer,

I am writing this letter about my best piece Vinnie. I like the story Vinnie because it is about a homeless person who finds a home. I shared my piece with Grahm and he said it was good. I feel good about my piece.

Sincerely,

"Vinnie" (Text 2B.2)

Vinnie was a normal kid, He dressed and looked the same as the other kids, Bur there was one thing that the other kids did not know about Vinnie: he was homeless. For the last two months Vinnie had been living in the toy store that closed down two years back. Vinnie had a job at the gas station, He only made four dollars an hour and he worked on the weekends from one to five. With the money he made he brought his clothes and food. He went to school on the weekdays.

Vinnie was running late that Friday, He was ten minutes late for school when he walked into math class. His teacher asked him, "Why are you late?"

Vinnie said, "I missed the bus."

When class was over his best friend Joe came up to Vinnie and asked him, "Do you want to come over to my house today?".

Vinnie said, "Sure."

So when class was over, they walked over to Joe's house. When they got there they had a snack. The two of them played for the rest of the day. Then Joe's mother mentioned that it was time for Vinnie to go home.

She said, "Should you call your mom?"

Vinnie said, "No, I will just walk."

Then Vinnie said, "Goodbye", and walked to the candy store which luckily, was only a few minutes away.

The next day, Joe came up to Vinnie in school and asked if they could go to Vinnie's house after school.

Vinnie said, "I don't think that would be a very good idea".

Joe said, "O.K., maybe we can do something tomorrow?"

"Yeah, see you tomorrow."

The next day Joe came up to Vinnie and asked him if he could come over. Vinnie said that he was grounded and he could not have any friends over. This went on for about one month, until one day Joe asked Vinnie why he could never do anything.

Vinnie said, "I don't want to talk about it right now".

Joe asked Vinnie if something was wrong.

Vinnie said, Yes, meet me in the park at 9:00p.m. tonight", and then he ran off.

Later that night, the two boys met in the park and Vinnie explained everything to Joe.

Joe asked Vinnie if he wanted to come live with him for a few weeks.

Vinnie said, "Sure". So that night Vinnie slept at Joe's house.

The next night the whole family went out to dinner, including Vinnie. When they were at the table waiting for their food Vinnie explained how he had become homeless and he had no mom or dad. After dinner they went home and went to bed.

The next day Vinnie had to go to work. When he was at work he purchased a lottery ticket for one dollar. The lottery for that week was one million dollars. After work Vinnie went home to Joe's house, and he gave the ticket to Joe's mom.

Later that week, every one was watching TV and the winning numbers for the lottery came on and Joe's mom checked to see if they won. They did not win, but they did get five of the six numbers right. They went down to the gas station and turned in their ticket. A few weeks later they received $300,000.00 in the mail.

Vinnie had been staying at Joe's for one month now and that night Joe's mom asked him if he wanted to live with them.

Vinnie said, "Sure".

After a few years Vinnie went off to college and now he is a lawyer.

The end

"Malachia" (Text 3B.2)

Chapter 1

It was a cold but amazingly clear day in Malachia. It was the kind of day that, when your alarm clock woke you up, you would curse at it, throw it at the wall, and then go back to sleep. Malachia was usually deserted, but today there was at least one thousand people out. The reason was that the mayor, Adam Hall, was to speak at one o'clock. He was to deliver, as he put it, "an historic announcement".

What the people of Malachia thought of it, well, let's just say they weren't pleased. They thought of Mayor Hall as, "the biggest lowlife that ever stepped on the face of this earth". And they were right. It was a miracle that Mayor Hall had been elected, and many said that he probably had been tinkering with the system. Many said he was the bigger loser than Gupta, while others disagreed. "Manu was not quite as bad as Bombard or Hall."

Malachia is the smallest town in the whole planet of Lafta, known before as the Parallel Earth. Some two hundred years before, the only planet with proven life, true Earth, had exploded. Luckily, most of the two hundred billion human lives had been spared.

Hall's speech was almost over. Most people did not throw tomatoes yet, but some could not resist. The speech was mostly, to everyone's dismay, about constructing more coal and oil refineries. Sure, it would give the unemployed jobs, but not many considering that the normal sized refinery only employed about 1,000 people.

One person, Vance Edwards Orr, put it nicely, "What exactly is the big deal with this coal and oil system? I think that it we didn't even touch the stuff, the other Earth never would have blown up and we never would be here on this damned planet." This was his subtle way of saying, "I hate this stupid place and I want to go back to where my ancestors lived and it's all the fault of coal and oil."

"Let it go. You may be right, it was the fault of coal and oil, but it was other things, too." Will Sheffer, Vance's best friend said soothingly.

"I'm not stopped for one minute. I am going to find out what happened, how it happened and why everyone is trying to cover it up."

"Maybe you are right, Why wouldn't they tell us what really happened? What if millions, even billions, of people died and they are trying to hide it? You're definitely right, but we will need more people to believe us, We'll have to get started real soon."

Vance paused, "Wait a minute! Hold on! We're moving a bit too fast here. Maybe we shouldn't do this, " he exclaimed.

That's weird, just a minute ago you wanted to find out really bad what was happening and now you're saying that maybe we shouldn't do it?"

"Well, I just think we are getting in a little too deep. I mean, it sounds like a Secret Service for God's sakes! All I wanted to have happen was for me to do this myself. It's my own personal matter."

Will then said, "But I wanted to know also and I'm sure other people do, too. So why don't you let us work as a group instead of you as one person. I think we would find out a lot more, because they'd think that more people care about it than just one person."

"Well, I don't know," Vance thought out loud, "I think I care more than most of the people but you might be right that we'd find out more. So I guess I'll work with all the people that we can round up."

"Good! Let's find some people,", exclaimed Will, happy that he convinced his friend.