ASSESSING WRITING 6(2) 1999 pp. 145–178
ISSN 1075-2935
PII S1075-2935(00)00010-6

Investigating Rater/Prompt Interactions in Writing Assessment: Quantitative and Qualitative Approaches

SARA CUSHING WEIGLE
Georgia State University

This study investigates how experienced and inexperienced raters score essays written by ESL students on two different prompts. The quantitative analysis using multi-faceted Rasch measurement, which provides measurements of rater severity and consistency, showed that the inexperienced raters were more severe than the experienced raters on one prompt but not on the other prompt, and that differences between the two groups of raters were eliminated following rater training. The qualitative analysis, which consisted of analysis of raters' think-aloud protocols while scoring essays, provided insights into reasons for these differences. Differences were related to the ease with which the scoring rubric could be applied to the two prompts and to differences in how the two groups of raters perceived the appropriateness of the prompts.
© 2000 Elsevier Science Inc.

INTRODUCTION

One of the most challenging issues in writing assessment is determining to what extent and under what circumstances different prompts can be considered equivalent. In direct tests of writing that involve timed impromptu essays, this issue is important for several reasons. Since examinees generally write on only one, or at most two different writing prompts, it is essential that the prompt represent as far as possible the domain of writing that is of interest. If examinees are offered a choice of prompts, it is important that they have equal opportunities to perform well on any of the choices. Finally, if different prompts are administered on different occasions, it is important that the prompts are of

Direct all correspondence to: S.C. Weigle, Georgia State University, Department of Applied Linguistics and ESL, P.O. Box 4099, Atlanta, Georgia 30302-4099.

approximately equal difficulty so that scores are not affected by the particular prompt that happens to have been administered.

Defining what is meant by prompt difficulty is not as straightforward as it may seem. Prompt difficulty is usually thought of in terms of scores: that is, prompts that elicit lower scoring essays are considered more difficult than those that elicit essays receiving higher scores. In a review of literature in L1 writing, Huot (1990a) discusses a number of prompt factors that affect scores on writing tests including, among others, discourse mode or purpose for writing (Quellmalz, Capell, & Chou, 1982; Hoetker, 1982), topic structure and wording (Smith et al., 1985), and degree of rhetorical specification (Brossell, 1983).

It is not always clear whether differences in scores are due to actual qualitative differences in the texts produced by test takers or to some aspect of the scoring process (i.e., the interaction of rates with essays written on particular prompts or prompt types). Some researchers have found qualitative differences in the texts. Crowhurst (1980) looked at narratives and argumentative writing across three grade levels and found that argumentative essays elicited significantly longer T-units than narratives. Furthermore, for tenth and twelfth graders, although not for sixth graders, length of T-unit was significantly related to holistic scores in the argument essays but not in the narratives. In a study of ESL student writers, Reid (1990) found significant differences between two essay types in essay length and lexical variables such as word length and percentage of content words, but no syntactic differences.

On the other hand, evidence is mounting that the scoring process itself is an important mitigating variable that can influence whether test scores differ across different discourse modes or prompt types. Hake (1986) found that essays that were pure narratives of personal experience were misgraded much more frequently than were expository essays using personal narration to illustrate or support an assertion. Purves (1992), reporting on a large-scale international study of school writing, notes that correlations across different functional essay types vary from country to country, and hypothesizes that these differences may be due to differences in raters rather than in actual student ability. In a study that is particularly relevant to the issue of the interaction of raters with different prompt types, Hamp-Lyons and Mathias (1994) found that essay topics that were judged more difficult by composition specialists tended to get higher scores than those judged to be easier, and suggested that raters may be unconsciously rewarding test takers who choose the more difficult prompt or may have lower expectations for that topic.

It is thus becoming clear that in addition to factors within the prompts themselves, rater variables have an important influence on

composition scores. One aspect of rater behavior that has been shown to influence test scores is rater expectations. Stock and Robinson (1987) go so far as to say that expectations may be as important as the quality of the text itself in determining composition scores. Diederich (1974) found that raters gave higher scores to the same essays when they were told that the essays were written by honors students than when they were told the essays were written by average students. Recent studies of handwritten versus word processed essays have demonstrated that he raters tend to score handwritten essays higher, in part because the expectations of formatting, grammatical, and spelling accuracy are higher in word processed essays and errors are thus more noticeable and glaring in these essays (Powers et al., 1994).

Another relevant rater characteristic that has been shown to affect tests scores is rater background. Several studies have found differences between novice and expert raters in terms of overall severity of grading and in terms of their rating strategies (Huot, 1988; Cumming, 1990; Shohamy, Gordon, & Kraemer, 1992; Weigle, 1994a). In second language writing, comparisons between ESL specialists and other raters (e.g., English faculty or other content area faculty) have demonstrated that raters from different disciplines apply different criteria to nonnative English writing (Mendelsohn & Cumming, 1987; Santos, 1988; Brown, 1991; Sweedler-Brown, 1993), highlighting the role of rater background experience in assigning scores to compositions.

Rater training is frequently cited as a means for compensating for different rater backgrounds and adjusting rater expectations so that variability in scoring based on disparate expectations is reduced (Charney, 1984; Huot, 1990b). While the results of training have seldom been demonstrated empirically, Freedman and Calfee (1983) found that slight differences in training had significant effects on overall scores, and Weigle (1994b) found that inexperienced raters did modify their expectations of student writing as a result of training.

To summarize, it is clear that prompt difficulty involves a complex set of interactions between student ability, aspects of the prompt itself, the scoring criteria as embodied in a scoring rubric and/or sample papers, and various rater characteristics, particularly rater background and the expectations that raters have for performance on a given prompt. While it may be impossible to completely tease apart the effects of these different variables, some light may be shed on them by the application of two methodologies that have arrived fairly recently on the scene of writing assessment: one quantitative—FACETS, or multifaceted Rasch measurement (Linacre, 1989); and one qualitative—verbal protocol analysis. The study described in this paper uses a combination of these methodologies to investigate differences in rater

perceptions of two different prompts and the relationship of these perceptions to actual scores given. Before discussing the study itself, it may be useful to provide a brief introduction to these methodologies.

MULTIFACETED RASCH MEASUREMENT (FACETS)

Multifaceted Rasch measurement is an application of Item Response Theory (for an introduction to IRT see Hambleton, Swaminathan, & Rogers, 1991), useful for analyzing data from any kind of examination that is scored by human raters, such as oral interviews or essay examinations, because it can provide estimates of the severity of raters, the difficulty of particular tasks, and other factors that influence ratings of ability. The theory underlying this measurement model is that the probability of a certain performance being given a particular score (for example, a four on a six-point scale) can be seen as a function of the examinee's ability and several other facets of the scoring situation, such as the severity of the rater, the difficult of the task, and the threshold of difficulty between the points on the scale (for example, between a four and a five). By estimating these other facets it is possible to obtain a more accurate estimate of the examinee's ability in the skill being tested.

The software package FACETS (Linacre & Wright, 1989-1994) has been developed to implement the multifaceted Rasch model. FACETS has been used to investigate rater behavior in performance assessment in a number of fields, such as acting ability (Myford, 1991a, 1991b) and medical certification examinations (Lunz, Wright & Linacre, 1990; Lunz, Stahl, Wright & Linacre, 1989). FACETS has also become fairly well known in language testing (McNamera & Adams, 1991; Stansfield & Kenyon, 1992; Bachman, Lynch, & Mason, 1995; Weigle, 1998) and a thorough discussion of the theory and its applications for language testing, particularly for tests of speaking and writing, can be found in McNamara (1996). The potential of FACETS for use in large-scale writing assessment has also recently been demonstrated by Englehard (1992, 1996) and Du et al. (1997).

FACETS is particularly useful as a means for examining the behavior of individual raters. The program provides several important statistics for each element of each facet (e.g., the ability of each examinee or the severity of each rater): a logit measure, its standard error, and model fit statistics. In the case of raters, the logit measure is an indication of the severity of each rater. The fit statistics indicate the degree to which each rater's ordering of examinees is consistent with the estimated ability of the examinees (Lunz, Stahl, Wright, & Linacre, 1989). Thus FACETS addresses two

important questions about raters: to what extent do individual raters differ from each other in terms of tendencies to grade harshly or leniently (i.e., rater severity); and, to what extent do specific raters differ from other raters in terms of their rank ordering of examinees (i.e., rater consistency)? Furthermore, as will be demonstrated in this paper, FACETS can be used to investigate differences among raters in terms of severity and consistency across different writing tasks. One strength of FACETS is that a fully crossed design is not necessary—that is, the program can estimate rater severity and examinee ability without requiring that all raters read all essays, as long as there is sufficient overlap among raters and essays.

VERBAL PROTOCOL ANALYSIS IN WRITING ASSESSMENT RESEARCH

Verbal protocol analysis involves training subjects to verbalize their thoughts as they perform a complex cognitive activity, as a means for gaining insight into otherwise covert mental processes. The analysis of verbal protocols has a long history in psychological research, but it was only with the work of Ericsson and Simon (1980, 1984) that a theory of verbal reports and methodology for collecting and analyzing protocol data were systematized. Much of the work in protocol analysis deals with problem solving, mathematics, or decision making: however, in recent years protocol analysis has also been applied to the study of language-related academic tasks such as composing processes (e.g., Hayes & Flower, 1980; Flower & Hayes, 1980) and test-taking (e.g., Cohen, 1984). While protocol analysis has been sometimes criticized for being subjective and providing inaccurate or inconclusive data (e.g., Nisbett & Wilson, 1977), Smagorinsky (1989) concludes from a review of the literature that protocol analysis can provide valuable data if the protocols are collected and analyzed in a systematic, theory-driven way and the analysis is substantiated by other forms of evidence.

Several researchers have used this methodology in both first and second language writing assessment research in an effort to discern what raters focus on when scoring writing (Vaughan, 1992; Connor & Carrell, 1993) and to investigate differences between novice and expert raters (Huot, 1988; Pula & Huot, 1993; Cumming, 1990; Weigle, 1994a, 1994b). While research in this area is still limited, verbal protocol analysis of rater behavior shows promise in getting "behind the curtain" of essay rating (Connor-Linton, 1995).

CONTEXT OF THE STUDY

The data for this study were collected as part of a larger study on the effects of training on raters of ESL compositions (Weigle, 1994a; see also

Weigle, 1994b, Weigle, 1998). The setting for this study is the composition subtest of an institutional ESL placement test at a large public university, the English as a Second Language Placement Examination (ESLPE), which also includes listening, reading, and at the time the data were collected, grammar subtests. The ESLPE composition subtest consists of a 50-minute essay on the student's choice between two prompts. One prompt requires students to interpret graphical information and make predictions based on this information (the Graph Interpretation prompt, hereinafter referred to as "Graph"), while the other prompt requires students to make and defend a choice based on information contained in a chart or table (the Choice Justification prompt, hereinafter referred to as "Choice"). At the time that this study was carried out, three Graph prompts and three Choice prompts were used in rotation on different forms of the ESLPE, but each form contained one Graph and one Choice prompt.

The compositions are rated using the ESLPE Rating Scale, which consists of three subscales (Content, Rhetorical Control, and Language). Each scale is divided into five two-point bands with descriptors for each band (see Appendix A)[1]. The total score is derived by doubling the language score and then summing all three subtest scores. Each essay is read by two raters, and the two scores are averaged. In cases of extreme score differences (five or more points), the essay is given to a third reader, and the two scores which are closest to each other are used in determining the final score.

The compositions are rated primarily by ESL faculty and teaching assistants (TAs), although occasionally other raters are hired with little or no experience in ESL teaching. All raters attend mandatory composition rater training, or "norming" sessions, which are led by the composition supervisor. The most extensive norming sessions, each lasting up to two hours, take place during fall quarter, when many new raters must be hired. In these norming sessions, raters discuss the prompts and the rating scale, read essays that represent the various scoring bands on the scale, and practice rating essays.

RESEARCH QUESTIONS

The research questions for the study are as follows:

1. To what extent do inexperienced and experienced raters differ in their ratings of the two prompts (Choice and Graph) in terms of severity and consistency before and after training?
2. What evidence from raters' verbal protocols and other qualitative data can be brought to bear on these differences?

SUBJECTS

The subjects for the study consisted of 16 people, eight from each of the following groups:

A. Raters who had never rated compositions at the institution and thus had never been exposed to the ESLPE composition prompts, student essays, or scoring guide and procedures used in scoring the ESLPE (hereinafter: new). These subjects included seven new ESL TAs with zero to ten years of teaching experience, and one non-TA graduate student in Applied Linguistics. Two of these raters had experience with composition rating, but not with the ESLPE. All eight new raters were female, and all were native speakers of English, with two exceptions: the non-TA was a native speaker of Navajo and another new rater was a native speaker of Korean, but both had learned English as young children and had native-like proficiency.

B. Experienced university ESL teachers who had rated ESLPE compositions before and thus were quite familiar with the rating scale, the composition prompts, and the level of student writing commonly found at the university. This group consisted of returning TAs, most of whom had at least two years prior experience with the rating scale and from two to ten years of teaching experience. Six of these raters were male and two were female. All were native speakers of English.

MATERIALS

The materials for the study included the following:

ESLPE Composition: A stratified random sample of sixty compositions from an earlier administration of the ESLPE was used. These were evenly divided between two composition prompts (one Choice and one Graph), found in Appendix B. The compositions were selected to represent all levels of proficiency among ESLPE examinees, based on the original scores assigned to each composition. To ensure that the two sets of compositions were written by students of approximately the same ability levels, the mean scores of both groups of examinees on the other sections of the ESLPE were compared by means of t-tests, using writing prompt as the independent variable and ESLPE scores on the combined listening/reading sections and grammar sections as the dependent variables. Results of these tests are found in Table 1.

As the table shows, students who wrote on the Graph prompt scored somewhat higher on the rest of the ESLPE than those who wrote on the Choice prompt, although these differences were not statistically sig-

Table 1. ESLPE Subscale Scores by Prompt

		Listening/Reading		Grammar	
Prompt	n	x	s.d.	x	s.d.
Choice	30	48.33	11.51	15.57	3.70
Graph	30	50.20	11.56	15.70	3.32

Listening/Reading: t = .5594, df = 58, p = .5780; Grammar: t = .1468, df = 58, p = .8838

nificant. While this is not conclusive evidence that the two groups are roughly equivalent in writing ability, it is at least a fairly adequate indication that the two groups are roughly equivalent in terms of overall English ability as measured by the other sections of the ESLPE.

ESLPE Rating Scale: The ESLPE Rating Scale, as described above, was used for all composition rating.

DATA COLLECTION

Pre-Training Data Collection

Step 1. Before the operational rater training session for the fall ESLPE, each subject was interviewed about relevant previous experience (teaching, composition rating, experience with nonnative speakers of English). All interviews were audiotaped.

Step 2. Subjects were given the ESLPE Rating Scale and 13 compositions on one topic (Choice or Graph) and asked to rate the compositions using this scoring guide. The design was counterbalanced so that half of the subjects read the Graph essays first and the Choice essays second, while the other half read the Choice essays first and the Graph essays second. A rating grid for each set of essays (Choice and Graph) was created which assigned compositions to raters to assure that each composition was read by approximately the same number of raters, that raters read compositions that spanned the range of abilities, and that enough overlap in rater/composition combinations was present for the quantitative analysis (specifically, making sure that there were no subsets of compositions which were only rated by a certain group of raters, who only read those compositions).

Step 3. After completing these ratings, subjects were given training in think-aloud procedures (Ericsson & Simon, 1984), and rated two additional compositions on the same prompt and two on the other prompt while verbalizing their thoughts into a tape recorder. These four compositions were the same for all raters. In most

cases these ratings were done in the presence of the researcher; however, occasionally the researcher was absent for brief periods while the think-aloud ratings were being done.

Step 4. Subjects were given 13 additional compositions on the second prompt to rate independently, using the ESLPE Rating Scale. They were instructed to rate the compositions as soon as possible, but in any case before the official norming sessions for the fall ESLPE administration took place. All raters complied with this instruction.

Data Collection During Training (Norming) Sessions

Step 1. All subjects participated in one of two of the ESLPE norming sessions, which lasted approximately 90 minutes. These norming sessions gave an introduction to the appropriate procedures for rating the ESLPE compositions, and provided opportunities for raters to see and discuss the scores given to essays in previous administrations. The norming sessions were videotaped.

Step 2. Subjects also participated in the operational rating session of the ESLPE compositions, but the data from these rating sessions were not used for this study. The ESLPE was administered five times over the period of ten days, and composition rating took place the day of the exam and on the following day. Subjects reported having spent from six to ten hours rating compositions over the course of the ESLPE administration and scoring sessions.

Post Training Data Collection

Step 1. Within the first two weeks after the operational rating sessions were completed, the researcher met with subjects individually and interviewed them regarding their experiences in the operational rating.

Step 2. Subjects were retrained in think-aloud procedures, and were then given six essays (three Choice and three Graph) and asked to rate them using the ESLPE Rating Scale while speaking their thoughts into the tape recorder. These six essays included the four that subjects had previously rated using think-aloud and one additional essay from each of the two forms, which the subjects had not rated previously. Subjects were told that they might have rated some of the essays before, and were asked to indicate on the scoring sheet for each essay whether they had rated it before, and if so, whether they could recall the scores they had given the essay.

Step 3. Subjects were given 26 additional essays (13 Choice and 13 Graph) to rate independently, and were asked to read the essays within a week. Most subjects complied with this request, although a few subjects took up to three weeks to finish the rating.

The essay rating scheme is summarized in Table 2.

Table 2. Summary of Essay Rating

Time	Silent rating (different, overlapping sets of essays for each rater)	Think-aloud (same essays for all raters)	Total
Pre-training	26 (13 each Choice & Graph)	4 (2 each Choice & Graph)	30
Post-training	26 (13 each Choice & Graph)	6 (3 each Choice & Graph)*	32

*Includes the four essays from the pre-training rating.

DATA ANALYSIS

A combination of quantitative and qualitative approaches to the data were used. Rater behavior before and after training was as modeled using the IRT program FACETS, versions 2.6 and 2.7 (Linacre & Wright, 1992, 1993), as discussed above. FACETS analysis for the two prompts were run separately. Because a preliminary analysis of the data revealed that the two most extreme scores (1 and 10) were rarely given, the FACETS analysis was simplified by collapsing the end points of the scale (1/2 and 9/10), essentially creating an eight-point scale rather than a ten-point scale.[2]

Differences in rater behavior across the two prompts were then investigated qualitatively, through an analysis of the audiotaped interviews, the norming sessions, and the think-aloud protocols, following procedures for qualitative data analysis outline in Strauss (1987) and Miles and Huberman (1983). Essentially this involves working back and forth between the data and a conceptual framework which informs, and in turn is informed by, a coding scheme for the data. As major themes in the data emerge, they can be described verbally or summarized in concise visual displays.

The videotapes of the norming session were viewed several times, with occasional excerpts transcribed; however, the videotapes were not transcribed in their entirety. Think-aloud protocols were transcribed in full, including false starts and pauses. Pauses of five seconds or longer were timed and are indicated in the transcripts by a number in parentheses representing the number of seconds. For example, (5.) represents a five second pause. Shorter pauses are indicated by a period within parentheses: (.)

For the purposes of this paper, the qualitative analysis focused on instances where raters spontaneously mentioned aspects or expectations for each of the two writing prompts. In addition, instances where the raters focused on descriptors within the rating scale in relation to essays on the two topics were coded.

RESULTS

FACETS Analysis

Results of the pretraining FACETS analyses for the Choice and Graph prompts are found in Table 3. For each rater, the analysis provides a severity estimate (measure logit), the error associated with this estimate, and fit statistics. In the table, the raters are arranged from most to least severe, as indicated by the measure logit. For both sets of essays, the most severe rater was Rater NEW1 (severity = .69 for Choice and 1.21 for Graph) and the most lenient rater was Rater NEW6 (severity = −.61 for Choice and −.54 for Graph). It should be noted, however, that the severity estimates in Table 2 are meaningful only within the context of the individual analysis and are not comparable across prompts. That is, the analysis tells us that the difference in severity between Rater NEW1 and Rater OLD6 (.69 logits −.40 logits = .29 logits) on the

Table 3. PRE Raters Measurement Report: Choice and Graph Separately

	Choice					Graph			
		Model	Infit			Measure		Infit	
Raters	Measure logit	Error	MnSq	Std	Raters	logit	Model Error	MnSq	Std
NEW1	0.69	0.12	1.6	2	NEW1	1.21	0.13	2.2	3
OLD6	0.40	0.12	0.7	−1	NEW4	0.45	0.12	0.9	0
NEW4	0.28	0.12	0.8	0	NEW2	0.37	0.12	0.4	−3
NEW3	0.21	0.12	1.5	2	NEW7	0.32	0.12	2.1	4
NEW8	0.19	0.12	1.0	0	NEW8	0.15	0.12	0.6	−2
OLD5	0.18	0.12	1.0	0	OLD1	0.02	0.11	0.6	−2
NEW7	0.16	0.12	2.1	4	NEW5	−0.00	0.11	0.9	0
NEW2	0.15	0.12	0.9	0	NEW3	−0.00	0.12	1.5	1
OLD2	−0.00	0.12	1.2	0	OLD6	−0.13	0.12	0.8	−1
OLD4	−0.10	0.12	0.7	−1	OLD5	−0.16	0.11	0.9	0
OLD1	−0.19	0.13	0.7	−1	OLD3	−0.17	0.12	0.4	−3
OLD3	−0.25	0.12	0.5	−2	OLD4	−0.22	0.12	1.0	0
NEW5	−0.29	0.13	0.6	−2	OLD2	−0.34	0.12	0.7	−1
OLD8	−0.38	0.13	0.9	0	OLD8	−0.45	0.13	1.0	0
OLD7	−0.43	0.13	0.7	−1	OLD7	−0.51	0.12	0.9	0
NEW6	−0.61	0.13	0.4	−3	NEW6	−0.54	0.12	0.9	0
MEAN	−0.00	0.12	1.0	−0.5	MEAN	−0.00	0.12	1.0	−0.5
S.D.	0.33	0.00	0.4	2.0	S.D.	0.43	0.00	0.5	2.2

Choice: RMSE: 0.12; Adj S.D.: 0.31; Separation: 2.50; Reliability: 0.86; Fixed (all same) chi-square: 112.72; d.f.: 15; significance: .00.

Graph: RMSE: 0.12; Adj S.D.: 0.41; Separation: 3.44; Reliability: 0.92; Fixed (all same) chi-square: 184.82; d.f.: 15; significance: .00.

Choice essays is more than twice the difference between Rater OLD6 and Rater NEW4 (.40 logits −.28 logits 2 logits); however, the analysis does not say anything about whether Rater NEW4 was more severe on the Choice essays (severity = .28) than on the Graph essays (severity = .45), as the two analyses were run separately and thus two different severity scales were constructed.[3]

It is clear from the table that the raters differ in their severity estimates, particularly for the Graph essays, with severity estimates ranging from −.54 logits to 1.21 logits; however, what the severity estimates themselves do not tell us is whether the differences in severity are meaningful or not. The FACETS analysis provides three statistics which address this issue: these are the separation index, the reliability, and the fixed (all same) chi square, found at the bottom of the table. The separation index is the ratio of the corrected standard deviation (Adj. S.D.) of element measures (in this case, rater severity estimates) to the root mean-square estimation error (RMSE). If the raters were equally severe, the standard deviation of the rater severity estimates should be equal to or smaller than the mean estimation error of the entire data set, resulting in a separation index of 1.00 or less; however, the prompt separation index is 2.50 for the Choice essays and 3.44 for the Graph essays, indicating that the variance among raters is substantially more than the error of estimates, particularly for the Graph essays, thus indicating that the raters are not equally severe.

The reliability statistic provided by the FACETS analysis indicates the degree to which the analysis reliably distinguishes between different levels of difficulty among the elements of the facet (in this case, the different raters). It is important to stress that for this analysis a *low* reliability is desirable, since in an ideal situation different raters would be equally severe and the analysis would not be able to distinguish severe raters from more lenient ones reliably; in this case, however, the reliability is quite high (.86 for the Choice essays, and .92 for the Graph essays), indicating that the analysis is reliably separating raters into different levels of severity. Finally, the fixed chi-square tests the null hypothesis that all of the elements of the facet are equal. For both sets of essays, the chi-square values are significant at $p = .00$, indicating that the null hypothesis must be rejected; in other words, the raters are not equal in severity.

The FACETS analysis also provides two measures of fit, or consistency: the infit and the outfit. The infit is the weighted mean-squared residual (i.e., the average difference between actual scores and the estimated scores provided by the analysis), which is sensitive to unexpected responses near the point where decisions are being made, while

the outfit is the unweighted mean-squared residual and is sensitive to extreme scores. For the purposes of this paper, only the infit statistics will be reported (labeled MnSq on the table). Although there are no hard and fast rules for determining what degree of fit is acceptable, a number of researchers (e.g., Lunz & Stahl, 1990) have found the lower and upper limits of .5 and 1.5, respectively, for mean squares to be useful for practical purposes. Fit statistics 1.5 or greater indicate misfit, or too much unpredictability in raters' scores, while fit statistics of .5 or less indicate overfit, or not enough variation in scores.

In addition to the mean squares, FACETS provides standardized infit statistics, which have an expected mean of 0 and standard deviation of 1. These statistics are useful for comparing the elements of a facet with each other, as they show the degree of variability in individual raters' ratings relative to the amount of variability in the entire data set. Standardized fit statistics greater than 2 or less than −2 are generally signs of misfit or overfit, respectively.

Applying these standards to Table 3, we can see that three raters show some degree of misfit on both sets of essays (Raters NEW1 and NEW7, and to a lesser extent, Rater NEW3). These statistics indicate that these raters' ordering of examinees was not consistent with the estimated ability measures of the examinees, and that the scores that they gave were highly unpredictable. Raters NEW1 and NEW7 are particularly misfitting on the Graph essays, with standardized infit statistics of 3 and 4, respectively. In addition, three raters Raters, NEW2, OLD3, and OLD4, show evidence of overfit, or lack of variability in their scores in comparison to the rest of the raters, based on standardized infit statistics of −3: Rater NEW6 on the Choice essays, and Raters NEW2 and OLD6 on the Graph essays. Overfit is often an indication that raters are preferring the central categories in a scale rather than the endpoints.

There are two additional comments to be made about Table 3. First, an inspection of the table reveals that, while the five most severe raters, and indeed seven out of the eight most severe raters on the Graph prompt are new raters, the separation between old and new raters for the Choice essays is not as pronounced. Note in particular that among the most severe raters on the Choice prompt are two OLD raters, Raters OLD6 and OLD5. Thus there appears to be a difference between the experienced and new raters in terms of their severity, particularly on Graph essays. The other phenomenon of note is the fact that, apart from the raters at the extreme ends of the severity continuum (Rater NEW1 at the top and raters OLD8, OLD7, and NEW6 at the bottom), there seems to be no consistent pattern of leniency or severity among

raters across the two prompts. For some raters, then, it appears that severity or leniency may be a general trait, while for others it may be prompt dependent.[4] Similarly, rater consistency is not the same across the two prompts.

In order to test whether the difference in overall severity between the two groups of raters is statistically significant, a Mann-Whitney U test for each prompt was performed. The Mann-Whitney U test is a non-parametric test that uses rank orders rather than measure values and is more appropriate than a t-test for small samples that are not normally distributed (Hatch & Lazaraton, 1991). Results of these tests are found in Table 4.

As the table indicates, there is no significant difference between experienced and new raters in severity for the Choice prompt ($Z = -1.55$, $p = .12$); however, the new raters are significantly more severe on the Graph prompt than are the experienced raters ($Z = -2.31$, $p = .02$). The severity difference between experienced and new raters on the Graph prompt provides insight into the larger separation index for the Graph essays discussed above: the greater severity of the new raters increases the overall spread of rater severity estimates for the whole set of raters.

To summarize, there was no significant difference in severity between the experienced and new raters before training on the Choice essays; however, the new raters as a group were significantly more severe on the Graph essays than were the experienced raters. We will now turn to the analysis of the post-training data before discussing the implications and possible causes of these results.

Table 5 shows the raters' measurement report for the two prompts following training. As the table shows, the spread of rater severity estimates (separation = 2.15 for Choice, 2.75 for Graph) are lower than the corresponding estimates for the pretraining data (2.50 and 3.44, respectively); however, for both prompts these figures still represent significant differences among raters. As in the pretraining data, the spread of raters severity estimates for the Graph prompt is greater than that for the Choice prompt,

Table 4. Mann-Whitney U Test: PRE NEW vs. OLD, by Prompt

		Choice		Graph	
Rater Type	n	Σ Rank	Mean Rank	Σ Rank	Mean Rank
NEW	8	57	7.13	46	5.75
OLD	8	79	9.88	90	11.25

Choice: $Z = -1.55$, $p = .12$; Graph: $Z = -2.31$, $p = .02$.

Table 5. Post Raters Measurement Report: Choice and Graph Separate

	Choice					Graph			
	Measure	Model	Infit			Measure	Model	Infit	
Raters	logit	Error	MnSq	Std	Raters	logit	Error	MnSq	Std
NEW4	0.61	0.14	1.2	1	NEW8	0.92	0.13	0.9	0
NEW3	0.58	0.14	1.1	0	OLD1	0.51	0.14	0.9	0
OLD8	0.33	0.14	1.2	1	NEW3	0.40	0.13	1.4	1
OLD4	0.30	0.14	0.9	0	NEW7	0.27	0.13	1.2	0
NEW8	0.15	0.14	0.8	0	OLD3	0.27	0.13	0.6	−2
OLD3	0.12	0.14	0.6	−2	NEW4	0.22	0.13	2.1	4
NEW1	0.06	0.14	0.9	0	OLD8	0.11	0.13	1.2	0
OLD6	0.03	0.14	0.8	−1	OLD4	0.01	0.13	0.6	−2
OLD2	−0.01	0.14	1.5	2	OLD2	−0.03	0.14	1.3	1
OLD1	−0.05	0.14	0.8	0	NEW5	−0.08	0.14	0.7	−1
NEW6	−0.16	0.14	1.0	0	OLD5	−0.29	0.14	0.8	−1
NEW7	−0.18	0.14	1.0	0	NEW1	−0.35	0.14	1.2	1
OLD5	−0.30	0.15	1.0	0	NEW6	−0.35	0.13	0.8	−1
NEW5	−0.46	0.15	1.2	0	NEW2	−0.41	0.14	0.5	−3
NEW2	−0.49	0.14	0.8	−1	OLD7	−0.56	0.14	0.8	−1
OLD7	−0.54	0.15	1.0	0	OLD6	−0.56	0.14	0.6	−2
MEAN	−0.00	0.14	1.0	−0.4	MEAN	−0.00	0.14	1.0	−0.3
S.D.	0.34	0.00	0.40	1.9	S.D.	0.40	0.00	0.4	1.8

Choice: RMSE: 0.14; Adj S.D.: 0.31; Separation: 2.15; Reliability: 0.82; Fixed (all same) chi-square: 90.57; d.f.: 15; significance: .00.

Graph: RMSE: 0.12; Adj S.D.: 0.41; Separation: 2.75; Reliability: 0.88; Fixed (all same) chi-square: 135.84; d.f.: 15; significance: .00.

although the difference between the two is not as great. Still, it seems to be easier for raters to judge the Choice essay more similarly than the Graph essays, with or without training. Perhaps what is most striking about the table is the fact that there is no clear pattern of new *vs* old raters in terms of severity, nor do individual raters show similarity across prompts either.

In terms of rater consistency, only one rater (Rater NEW4) is misfitting in the Post Graph data (infit = 2.1, standardized infit = 4), indicating that her rank ordering of examinees on the Graph prompt was not consistent with that of the other raters. On the Choice essays, Rater OLD2 is the least consistent, with an infit of 1.5 and standardized infit of 2. Overall, the fit statistics for the entire sample of raters are much better than they were in the Pre-essays, indicating that the pattern of ratings is consistent enough for the analysis to model rater severity accurately.

Turning to a consideration of the two groups of raters, Table 6 shows the Mann-Whitney U tests for the post-training data. As the table

Table 6. Mann-Whitney U Test; POST NEW vs. OLD, by Prompt

		Choice		Graph	
Rater Type	n	Σ Rank	Mean Rank	Σ Rank	Mean Rank
NEW	8	67	8.38	63.5	7.94
OLD	8	69	8.63	72.5	9.06

Choice: $Z = -1.05$, $p = .92$; Graph: $Z = -.473$, $p = .64$.

shows, there are no significant differences between the two groups of raters on either prompt. Thus we can see that the difference in severity between the two groups has lessened considerably.

To summarize the results of the FACETS analyses, we have seen that the inexperienced raters tended to be more severe on the Graph essays than their more experienced counterparts before training, and that these differences were no longer present to the same extent following training, nor were they found to be as great a degree with the Choice essays. Before moving on to the possible causes of these differences, it might be informative to take a brief look at what effects these differences would have in operational testing. It will be recalled that in operational testing, the ESLPE composition score is calculated by doubling the language score and adding the result to the content and organization scores, for a maximum of 40 points, and that a difference between two raters of more than five points necessitates a third reading of the essay. As a way of looking at group differences in rating essays, the total score for each composition given by each rater was calculated, and an average score across all raters within rater groups was calculated. The raw scores for each composition given by all new raters before training were averaged and compared with the averaged raw scores given by experienced raters. Table 7 shows the number of essays whose average score from the two groups of raters differed by five points or more. As the table shows, ten out of 60 compositions, or about 17%, differed by at least five points between the average score given by experienced raters and that for inexperienced raters before training. The majority of these compositions (seven out of ten) were on the Graph topic. In all except one case (a Choice composition), the average score from the new raters was lower than that of the experienced raters. In the Post ratings, in contrast, only two of the essays, one on each prompt, show an average difference of at least five points between the two groups of raters. This finding once again reflects the overall greater severity of the new raters, particularly on the Graph essay, before training, and the reduction of differences in rating behavior by the two groups following training.

Table 7. Number of Compositions Rated on Average ≥ 5 Points Differently by New and Old Raters

	PRE			POST		
Score Difference	Choice	Graph	Total	Choice	Graph	Total
NEW>OLD	1	0	1	1	0	1
OLD>NEW	2	7	9	0	1	1
TOTAL	3	7	10	1	1	2

Discussion

The quantitative results suggest that the effects of training may be more important for some types of prompts than for others, as the differences between experienced and inexperienced raters were much more salient on the Graph essays than on the Choice essays; by themselves, however, the quantitative data do not tell us why this is so. It may be that the ESLPE scoring rubric is more easily and consistently applied to Choice essays than to Graph essays, since both groups of raters were equally sever on these essays before training. There may be something about the Graph essays that leads inexperienced raters to judge them more harshly than experienced raters do, or about the scoring rubric which lends itself to harsher interpretation for Graph essays than for Choice essays when it is used by inexperienced raters. Alternatively, the experienced raters, who have had experience rating essays on both prompts before, may be more lenient on the Graph essays for other reasons, such as overfamiliarity with the Choice prompt, which tends to be chosen with greater frequency than the Graph prompt in actual ESLPE administrations. We will now turn to the qualitative data to provide more insights into these issues.

QUALITATIVE ANALYSIS

The primary source of qualitative data for the study was the verbal protocols of raters rating the same four essays, two on each topic.[5] Additional qualitative data came from the videotaped norming sessions and rater interviews. Before moving to the analysis of the verbal protocols, I will briefly describe the discussion of the two different prompts in the two norming sessions. It will be recalled that there were two different norming sessions, each of which was attended by approximately half the raters.

The Choice prompt type was discussed at length in the first norming session, with several experienced raters expressing dissatisfaction with

these prompts. The "fatigue factor" was mentioned by the supervisor in both sessions, referring to the fact that these prompts tend to be chosen more than the Graph prompts and that it is easy for raters to become bored with the topic quickly. In the second norming session, comments about the Choice prompts were made primarily by the supervisor, not the raters, while in the first norming session both the supervisor and the experienced raters commented frequently about these prompts. The only comments made about the Graph prompts concerned the clarity of the Graph in the Population prompt, which was seen as confusing and hard to read by many raters. This issue came up in both norming sessions.

At no time in either session was the issue of prompt equivalence raised. That is, there was no attempt to make sure that raters were applying the same standards to both prompt types. This issue was only implicitly addressed in discussions about the highest rated essay in the Choice packet, when the supervisor stated that the rating team had felt that this particular essay was as thoughtful a response as one could hope for on this particular prompt. This comment implied that one could be more thoughtful on the Graph prompts, for which it was easier to give 9s and 10s; however, no attempt was made to compare the qualities of highly rated essays across prompt types to ensure that they were equivalent in quality.

Verbal Protocols: Effects of Prompt Differences

In this section I will focus more closely on one specific aspect of the quantitative results: the differences between the experienced and the new raters before training as they relate to the two different writing prompts. Of particular interest is why the new raters were significantly more severe on the Graph essays than were the experienced raters, while there were no such differences on the Choice essays. Given these results, we can speculate that there is something about the Graph essays themselves, or about the match of these essays to the descriptors on the scoring guide, which make them more likely than Choice essays to be misrated by untrained raters.

Because the essays for which both PRE and POST verbal protocols are available are limited to two essays per prompt, definitive statements about the causes of these differences cannot be made; however, some educated guesses are possible, which may provide impetus for further research on this topic. In this section I will present two areas of difference found in the verbal protocols of the experienced and new raters which may help to account for these results. First of all, I will discuss

the nature of the prompts in terms of the demands placed on writers and the rhetorical strategies typically used to respond to each prompt. I will relate the prompt differences to comments made by raters while rating the two topics and show ways in which the Graph essays, in particular, were problematic for new raters to rate before training. The second area of difference has to do with the raters' subjective reactions to the prompts themselves, which may have an impact on raters' scoring patterns.

One reason that new raters were not able to rate the Graph essays consistently before training has to do with the difference between the two prompts in terms of the demand of each task and the rhetorical devices which are typically employed in essays on each prompt. The Choice prompt used in this study is fairly straightforward. Examinees are given information about three different jobs in terms of salary, working hours per week, vacation, and job satisfaction, and are asked to state which job they would prefer any why. The examinee does not need to bring any outside information to the task of writing an essay on this topic except for his or her own preferences. This type of task lends itself very easily to variations on the traditional five-paragraph essay with, for example, an introduction stating the thesis (the job preferred), three body paragraphs dealing with each job or with the advantages and disadvantages of the preferred job, and a conclusion restating the examinee's choice. The vast majority of examinees use one of a very few rhetorical strategies in organizing a paper on this topic, and the organization is typically quite transparent. The personal nature of the topic is also such that examinees frequently adopt a conversational tone rather than a more formal, academic style.

The Graph prompt is structured somewhat differently and tends to elicit a different organizational strategy. The prompt asks three separate questions which require examinees to perform essentially three different tasks: to interpret the graph in terms of changes in the U.S. population, to enumerate some problems that will occur as a result of the changes, and to propose some solutions to these problems. Most students choose to organize their essay by answering each question in turn, rather than by stating an explicit thesis addressing all three of the questions. In terms of content, examinees must bring more world knowledge to this writing task than to the Choice prompt, as they need to be aware of the types of problems that may be associated with an aging population. This prompt also tends to elicit more formal language than the Choice topic due to its more academic nature.

Following this brief description of the demands of the two writing prompts, I will now consider evidence from the verbal protocols that the differences between the two prompts did make a difference in rater

behavior, particularly before training. I will focus on rhetorical control in this section, since a preliminary analysis of the transcripts revealed that the differences between experienced and new raters in rating the two composition topics were most pronounced in this area. The transcripts from the new raters rating the two Graph essays before training show that, in some cases, these raters found it difficult to relate the organization of essays on the Graph prompt to descriptors on the scoring guide, particularly in the area of Rhetorical Control. Rater NEW2 comments extensively on this difficulty:

(1)

N2PRGR9[6]: Okay, I'm just noticing too that in composition topic number 2 [Choice] there's more of a controlling idea to me that seems more clear than this first one [Graph]. I really don't know exactly what the controlling idea would be in the first one. Except that there's changes in the population. . . . . I don't know. This seems more difficult for me to organize, I guess. I'm trying to think, how would I organize that? What's a good way to organize an essay on this topic?

This rater comments on the difficulty of locating the "controlling idea" of the essay, a key concept in the Rhetorical Control subscale of the ESLPE scoring guide. The first descriptor in each of bands 3-4, 5-6, and 7-8 contains a statement about the effectiveness of the presentation of the controlling idea in the introduction and in the conclusion. In a Choice essay, it is generally easy to locate something that might be called a controlling idea, since nearly all such essays contain a statement of which job the examinee prefers. In the Graph essays, on the other hand, finding a controlling idea, or even defining what constitutes a controlling idea, is a somewhat more difficult task, since these essays tend not to have explicit thesis statements. Thus, for several of the new raters it may have been difficult to relate this particular phrase in the scoring guide to the Graph essays, which may have led to lower scores.

Another difficulty in rating the Graph essays for some of the new raters was in determining what constituted an appropriate introduction or conclusion to these essays. There was some discussion in both of the norming sessions about this issue, as the descriptors in the Rhetorical Control subscale on the scoring guide seem to place a great deal of emphasis on these parts of the essay. Introductions and conclusions seemed to be a problem more for the Graph essays than for the Choice essays. Essay GRAPH9, in particular, was problematic in this regard. This can be illustrated by comparing the comments made by raters about the introduction of this essay. Before training, six out of seven

new raters commented on the introduction while rating rhetorical control. Of these, two reacted positively while four reacted negatively. In contrast, six out of eight experienced raters had positive reactions to the introduction of this essay, while only one had a negative reaction and one did not comment at all. After training, six out of seven new raters and all eight experienced raters made positive comments about the introduction to this essay. Some of the raters' comments from the pretraining rating illustrate these differences of opinion[7]:

(2)

N2PRGR9: Introduction I think includes too much of this person's imposing their own opinions or what they consider to be their knowledge of the figures in the graph, so maybe it could have been a little more objective. (RC = 6)

N5PRGR9: I'm seeing what's in the introductory paragraph. The person's giving reasons for why people are living longer and why there is a decrease in young people. Now the person gets to the graph. Which is a little different than how others have done it. But it's a nice general introduction before specifically talking about the percentages in the graph. (RC = 6)

N6PRGR9: So they're trying to explain why population changed at the beginning. . . I think it's weird because the beginning paragraph isn't really an introduction, it's a totally different thing. (RC = 8)

N7PRGR9: Introduction. That's kind of out of the blue. (RC = 4)

O5PRGR9: Interesting, this person doesn't actually refer to the graph here but you know what he's talking about. . . there's an introduction that I really like. (RC = 8)

O6PRGR9: This person is explaining the changes in the graph without explicitly saying what these changes are, but that's pretty, it's better really this way, I think. . . I like that introduction, more sophisticated. (RC = 8)

O7PRGR9: Beginning's very general statements. Rather than the typical description of the data. It's introducing birth control. The cause of the changes. Which seems a little bit irrelevant because the task does not really require you to discuss the causes as far as I know. (RC = 6)

As these examples show, there was a variety of opinion as to whether the author of this paper had begun it appropriately. While positive and negative reactions came from both groups of raters (new and experienced), the general tendency was for new raters to find it irrelevant or off topic and for experienced raters to find it an unusual but interesting way to begin the paper. For those raters who reacted negatively to this

paper's introduction, it seemed to violate their expectations of what an appropriate introduction to this paper would be. Those raters who liked the introduction found it to be somewhat out of the ordinary, but tended to see that as a positive aspect of the paper.[8]

Most of the comments regarding the introduction to this essay contain a description of the opening paragraph and what the author is trying to do in the paragraph. Such descriptions are rarely found in raters' comments regarding the introduction to the Choice essays. For these essays, the comments made by both experienced and new raters tended to be more directly tied to the descriptors on the rating scale. The following quotes are from some of the same raters quoted in Example 2 above, this time from Essay CHC7. Direct quotes from the scoring guide are in italics.

(3)

N5PRCH7: Um, rhetorical, (.) um let's see. I'm looking at 9 or 10, because this seemed to be very easy to read when I read it. *Introduction and conclusion effectively fulfill their separate purposes: the introduction effectively orients the reader to the topic (.) separate, yet cohesive, sentences form a well-connected series of ideas or logical steps.*

N6PRCH7: And the rhetorical control, they make paragraphs but (.) I don't like—the introduction and conclusion seem very token, and the way they introduce paragraphs is extremely boring.

O6PRCH7: Um, *introduction presents the controlling idea mechanically,* yes, *conclusion doesn't give new insights,* no.

In these quotes, and indeed for both of the Choice essays, the raters tend to go directly to the scoring guide when considering the rhetorical control, rather than first going back to the essay to discern the structure, as in Example 2. It is possible that the transparent organization of the Choice essays makes it easier for raters to hold the structure of the essay in mind after a single reading, so that they can make their judgments more quickly. It is also possible that these essays are more directly applicable to the descriptors on the scoring guide, at least in the area of rhetorical control, than are the Graph essays, which require some interpretation before they can be matched to the descriptors. Although it is impossible to show definitively that the differences between the two prompts in rhetorical structure led directly to the more severe ratings on the part of the new raters as a group, particularly because of the limited sample of essays used for the think-aloud portion of this study, the examples cited here suggest that these differences may have led to differences in rating behavior. Future studies could look at this issue by comparing think-aloud protocols on a wider variety of essays on both topics.

Table 8. Positive and Negative Comments about the Writing Prompts (+ = Positive, − = Negative)

	Choice		Graph	
RATER	PRE	POST	PRE	POST
NEW1	−	−		
NEW2		+	−	−
NEW5	+			
NEW6				−
NEW7	+			
OLD1	−			
OLD3		−		
OLD4	−	−		
OLD5	−	−		
OLD6		−		
OLD7	−			

The second area to be considered in this section is the comments made by raters about the writing prompts themselves. These comments were made in the context of rating the essays and had to do with raters' opinions of the prompt in general, and in comparison to the other prompt. A number of raters expressed different reactions to the two writing prompts, summarized in Table 8, as either positive comments (symbolized by "+") or negative comments (symbolized by "–"). Examples of these comments are found in Examples 4 through 7 below.

The table shows that comments about the Choice prompt were made much more frequently than comments about the Graph prompt. The table also shows that, of the five new raters who made subjective comments about the two writing prompts before training, four of them seemed to prefer the Choice prompt, stating that it was easier to read, required less background knowledge, or was less confusing than the Graph topic. The fifth rater, Rater NEW1, felt that the Choice topic was not academic and was too personal. After training, Rater NEW1 made similar comments, but the only other new rater who made comments about the two topics was Rater NEW2, who still had negative comments about the Graph topic.

In contrast, three of the experienced raters made negative comments about the Choice prompt before training, and five did so after training; altogether, six out of the eight experienced raters made at least one negative comment about the Choice topic at one time or another. None of the experienced raters commented on the Graph prompt while rating the Graph essays. Example 4 shows comments from two of the new

raters who prefer the Choice prompt. Both of these raters felt that the Choice prompt was easier to write about and/or easier to rate.

(4)

N2PSCHC2: I think I like this one [Choice] better because people would find it easier to answer than the first one [Graph]. And the first one, you have to know about cultural background and and other areas, um, about the U.S. You have to do a lot more speculation. This one you can just kind of think about your own personal preference, and it's probably easier for students to answer.

N5PSCHC7: And I'm thinking that this topic [Choice] is much easier to read about, is that because it's really easier to answer or because I think it's easier to answer?

In contrast, Rater OLD3 saw the fact that the Choice prompt was easier as a disadvantage, as can be seen in Example 5.

(5)

O3PSCHC2: Um, the problem with this prompt is that it doesn't allow the students um much chance to do any better writing. Or better thinking than the other one. It's the this is the easy prompt, the easy one out. And everybody writes exactly the same thing, no matter which, I mean if they choose Job C all the Job C essays are the same, and all the Job A essays are the same. All the Job B essays are the same, because then they say well A would be wonderful because of the job satisfaction and C is too many hours and so therefore I choose Job B. I mean they all tend to write exactly the same essay and so that's kind of disappointing.

This rater sees the simplicity of the prompt as a hindrance to writers, because it does not allow them to display their best writing. In addition, this rater finds the similarity of the essays on this prompt to be somewhat frustrating, since there are so few ways to approach the topic.

The other major complaint against the Choice topic leveled by raters is that it is not academic. This is illustrated by Example 6. Similar comments were made by one new rater (Rater NEW1) and three of the experienced raters.

(6)

O4PRCHC7: Then again, the problem comes back. Is this prompt conducive to a good essay? Is it asking you a question that's well devel-

opable in a really well-planned academic essay? Always have had doubts about this prompt.

It is unclear to what extent the raters' views of the writing prompts actually affected the scores that they gave; however, the transcripts provide at least one clear example of a rater who changed a score in reaction to the prompt:

(7)

O5PRCHC7: Well, I mean the topic is so simple that I think this person covered everything. . . . But, uh, maybe I should have given it an 8 [rather than a 7], actually, Yeah, why not? I'll strike out at the task, it's not the person's fault that they chose this, I mean they were given the choice.

In this excerpt, the rater had decided to give a score of 7 for content on this essay, but changes his mind and raises the score to an 8, stating explicitly that he was doing so in order to "strike out at the task." While this is the only example in the verbal protocols of a rater actually changing a score in this way, it is likely that prompt considerations had an effect on scores in other, more subtle ways. Further research using verbal protocols of a larger number of essays might elucidate this issue.

One reason for the consistency of experienced raters in making negative comments about the Choice prompt may be the fact that this prompt tends to be chosen quite a bit more frequently than the Graph prompt in the operational examination. The reasons for this are unknown, but it is likely that the Choice prompt seems easier to many students, perhaps because the topic requires less background knowledge or it is visually easier to process than the Graph prompt.[9] Because of its greater frequency, and because of the limited number of approaches to the topic, the Choice prompt tends to be subject to the "fatigue factor," as discussed previously. Example 5 above illustrates this phenomenon. This, in turn, may have an effect on raters' scores, particularly if raters are reluctant to give high scores to essays on this prompt.

To summarize, in this section we have seen two separate but complementary factors in rating the two essay prompts which may well have led to differences in overall severity between the experienced and the new raters before training. The Graph essays tended to be more difficult to relate to the descriptors on the scoring guide, which may have caused the less experienced raters to judge them more harshly than the Choice essays, which were rhetorically simpler and possibly easier to judge for untrained raters. Furthermore, the variety of approaches taken to the Graph essays may have been easier for experienced raters to adjust to, as experi-

enced raters tend to have a larger reading repertoire upon which to draw when scoring essays (Huot, 1988). On the other hand, the personal nature of the Choice essays, their lack of academic focus, and the limited number of response possibilities caused problems for the more experienced raters, who may have been suffering from overfamiliarity with these essays.

DISCUSSION/CONCLUSION

The study described in this paper applied both quantitative and qualitative methodologies to address, albeit in a limited way, one of the most vexing problems in writing assessment for both L1 and L2 contexts: the issue of task equivalence. The quantitative results shoed that raters did in fact treat the two prompts differently, particularly before training, while the qualitative results provided insights into the reasons for these differences in rater behavior that would have been impossible to discover otherwise. To be sure, the study is limited to one particular set of circumstances at one university, and thus the results obtained from the study may not be generalizable outside of those circumstances. From the data presented here we may not be able to come to conclusions about rater behavior in general, the differences in rater severity between experienced and inexperienced raters, or the difficulty of specific writing task types; however, the study does have important implications of the results for writing assessment in both first and second languages. First of all, the study demonstrates that important differences in rater behavior across different writing tasks may not be accessible through traditional methods such as comparing mean scores. Global measures such as mean scores may mask subtle task effects in scoring that are only revealed through a quantitative approach such as FACETS or a qualitative approach such as verbal protocol analysis.

Second, the results highlight the importance, but also the limitations of rater training, in getting raters to agree on the quality of student writing. For some task types, rater training may be more essential than for other task types. As assessment specialists attempt to measure writing in more authentic, meaningful ways, necessitating the use of new task types that may be unfamiliar to raters, more training and discussion among raters may be needed to reach consensus on how to judge these less traditional writing samples; the FACETS results from this study also serve as a reminder, however, that even on the most prosaic of topics with the most predictable outcomes, raters will not necessarily agree completely on scores. Some raters will be more severe than others, no matter how much training they are subjected to. Tools such as FACETS can be useful in taking rater severity into account when reporting scores, or at the very least score users need to be reminded that

essay rating is still a subjective process and scores should be interpreted in light of this fact.

The qualitative results of the study, in particular, underscore the complexity of the relationship between rater background, the scoring rubric, the prompt, and rater training in writing assessment. A scoring rubric that is developed with one particular prompt type in mind may not necessarily be appropriate for another prompt type, and this can be an argument either for limiting the prompt types to a very restricted range or for developing primary trait scoring guides for different prompt types (Lloyd-Jones, 1977; see also Hamp-Lyons, 1991, on developing multiple-trait scoring guides for ESL contexts). Furthermore, differences in expectations of performance on different prompts should be raised as an issue in rater training so that raters may learn to compensate for any unconscious bias in favor of one prompt versus another.

Finally, perhaps the most important implication of the study is as an illustration of how quantitative and qualitative analyses can complement each other. The quantitative results demonstrated that the groups of raters were rating the two essay types differently, but the qualitative results were equally important in providing insights into the causes of these differences. While applying both methodologies can be time consuming, the payoff in terms of our understanding of the process of writing assessment, and ultimately in improving the validity of our assessments, may be worth the effort.

NOTES

1. The rating score for the ESLPE has been substantially revised since this study was carried out.
2. Additional technical details about the FACETS analysis can be found in Weigle 1994a.
3. A direct comparison of rater severity across the two prompts would require that the same examinees write essays on both prompts, which was not the case for this study.
4. FACETS is capable of detecting bias in raters (that is, when raters are systemically more or less severe on some tasks versus others) but this analysis would require the same examinees to write on both prompts, which was not the case for this study.
5. As noted in the Data Collection section, raters provided think-aloud protocols for six essays in the post-training rating sessions; however, for the purposes of this paper only the protocols for the four essays that had also been read before training will be discussed.
6. Examples are labelled as follows: Rater (e.g., N2 = Rater NEW2), Time (NRM = norming session interview, PR = PRE, PS = POST), Composition (e.g., GR9).
7. The score given to this essay on rhetorical control by each rater is found in parentheses at the end of the excerpt.
8. It should be noted that there is not a direct correspondence between the rater's opinion of the essay's introduction and the Rhetorical Control score given, since the introduction is just one factor to be considered in assigning this score.

9. The reasons for choosing a particular essay prompt is a very important but under-researched topic. One recent study (Polio and Glew, 1996) suggests that students choose topics for a variety of reasons, one of the most frequently cited of which is familiarity with the topic.

Acknowledgments: The author would like to thank Patricia Carrell, Brian Huot, Maria Ines Valsecchi, Cynthia Walker, and two anonymous *Writing Assessment* reviewers for their helpful comments on earlier versions of this manuscript.

REFERENCES

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12* (2), 238–57.

Brossell, G. (1983). Rhetorical specification in essay topics. *College English, 45*, 165–73.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly, 25* (4), 587–603.

Connor, U., & Carrell, P. (1993). The interpretation of tasks by writers and readers in holistically rated direct assessment of writing. In J. G. Carson and I. Leki, *Reading in the composition classroom: Second language perspectives* (pp. 141–160). Boston, MA: Heinle & Heinle.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing. *Research in the teaching of English, 18*, 65–81.

Cohen, A. D. (1984). On taking language tests: What the students report. *Language Testing, 1* (1), 70–81.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly, 29*, 762–65.

Crowhurst, M. (1980). Syntactic complexity and teachers' ratings of narrations and arguments. *Research in the Teaching of English, 13*, 223–31.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*, 31–51.

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

Du, Y., Wright, B., & Brown, W. L. (1997). *Raters and single prompt-to-prompt equating using the Facets model in a writing performance assessment.* Paper presented at the International Objective Measurement Conference, Chicago, IL. Eric Document #ED410291.

Engelhard, G. (1992). The measurement of writing with a many-faceted Rasch model. *Applied Measurement in Education, 5* (3), 71–91.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33* (1), 56–70.

Ericsson, K. A., & Simon, H. (1980). Verbal reports as data. *Psychological Review, 87*, 215–251.

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

Flower, L. S., & Hayes, J. R. (1980). The dynamics of composing: Making plans and jug-

gling constraints. In L. W. Gregg & E. R. Steinberg, (Eds.), *Cognitive processes in writing,* (pp. 31–50). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

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

Hamp-Lyons, L., & Mathias, S. P. (1994). Examining experts' judgements of task difficulty of essay tests. *Journal of Second Language Writing, 3,* 49–68.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics.* New York: Newbury House Publishers.

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hoetker, J. (1982). Essay examination topics and student writing. *College Composition and Communication, 33,* 377–92.

Huot, B. (1988). The validity of holistic scoring: *A comparison of the talk-aloud protocols of expert and novice holistic raters.* Unpublished manuscript.

Huot, B. (1990a). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60,* 237–63.

Huot, B. (1990b). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41,* 201–13.

Kroll, B., & Reid, J. (1994). Guidelines for designing writing prompts: clarifications, caveats and cautions. *Journal of Second Language Writing, 3* (3), 231–55.

Linacre, J. M. (1989). *Many-faceted Rasch measurement.* Chicago, IL: MESA Press.

Linacre, J.M. & Wright, B.D. (1992). *A user's guide to FACETS (Version 2.6).* Chicago, IL: MESA Press.

Linacre, J. M., & Wright, B. D. (1993). *A user's guide to FACETS (Version 2.7).* Chicago, IL: MESA Press.

Lloyd-Jones, R. (1977). Primary trait scoring. In C.R. Cooper & L. Odell (eds.), *Evaluating writing: Describing, measuring, judging* (pp. 33–66). Urbana, IL: NCTE.

Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions, 13,* 425–44.

Lunz, M. E., Stahl, J. A., Wright, B. D., & Linacre, J. M. (1989). *Variation among examiners and protocols on oral examinations.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3* (4), 331–45.

McNamara, T. F. (1996). *Measuring second language performance.* New York: Addison Wesley Longman Limited.

McNamara, T. F., & Adams, R. J. (1991). *Exploring rater behavior with Rasch techniques.* Paper presented at the Language Testing Research Colloquium, Princeton, NJ.

Mendelsohn, D., & Cumming, A. (1987). Professors' ratings of language use and rhetorical organization in ESL compositions. *TESL Canada Journal, 5* (1), 9–26.

Miles, M., & Huberman, M. (1983). *Qualitative data analysis.* Beverly Hills, CA: Sage Publications.

Myford, C. M. (1991a). *Assessment of acting ability.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Myford, C. M. (1991b). *Judging acting ability: the transition from the novice to expert.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Nisbett, R. E., & Wilson, W. T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–59.

Polio, C., & Glew, M. (1996). ESL writing assessment prompts: How students choose. *Journal of Second Language Writing, 5* (1), 35–49.

Powers, D., Fowles, M., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word processed essays. *Journal of Educational Measurement, 31* (3), 220–33.

Pula, J. J., & Huot, B. A. (1993). A model of background influences on holistic raters. In M. M. Williamson and B. A. Huot (Eds.), *Validating holistic scoring for writing assessment. Theoretical and empirical foundations* (pp. 237–65). Cresskill, NJ: Hampton Press, Inc.

Purves, A. (1992). Reflection on research and assessment in written composition. *Research in the Teaching of English, 26*, 108–22.

Quellmalz, E. S., Capell, F., & Chou, C. P. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement, 19*, 241–58.

Reid, J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 191–210). Cambridge: Cambridge University Press.

Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly, 22* (1), 69–90.

Shohamy, E., Gordon, C., & Kramer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal, 76* (1), 27–33.

Smagorinsky, P. (1989). The reliability and validity of protocol analysis. *Written Communication, 6* (4), 463–79.

Smith, W.L., Hull, G.A., Land, R.E., Moore, M.T., Ball, C., Dunham, D.E., Hickey, L.S., & Ruzich, C.W. (1985). Some effects of varying the structure of a topic on college students' writing. *Written Communication, 2*, 73–89.

Stansfield, C. W., & Kenyon, D. M. (1992). *Comparing the scaling of speaking tasks by language teachers and by ACTFL guidelines.* Paper presented ta the Language Testing Research Colloquium, Vancouver, B.C.

Stock, P. L., & Robinson, J. L. (1987). Taking on testing: Teachers as researchers. *English Education, 19*, 93–121.

Strauss, A. (1987). *Qualitative analysis for social scientists.* Cambridge: Cambridge University Press.

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

Vaughan, C. (1992). Holistic assessment: What goes on in the rater's mind? In L. Hamp-

Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–26). Norwood, NJ: Ablex.

Weigle, S. (1994a). *Effects of training on raters of ESL compositions: Quantitative and qualitative approaches.* PhD dissertation, University of California, Los Angeles.

Weigle, S. (1994b). Effects of training on raters of ESL compositions. *Language Testing, 11*, 197–223.

Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing, 15* (2), 263–87.

APPENDIX A: ESLPE RATING SCALE

Content

9-10

a. The essay fulfills the assignment well and treats the topic with sophistication. The main idea is clear.
b. Support is relevant, thorough, and credible.

7-8

a. The essay addresses the assignment appropriately* and is well-developed. The main idea is clear.
b. Most of the arguments/ideas are well supported.

5-6

a. The essay addresses the topic appropriately, but may not be well-developed. OR The essay only addresses part of the topic, but develops that part sufficiently.
b. Some statements may not be supported or unrelated to main idea.

3-4

a. The essay is inappropriate to assigned topic OR the main idea is not evident.
b. The essay contains unsupported or irrelevant statements.

1-2

a. The paper lacks a clear main idea.
b. Several statements are unsupported, and ideas are not developed.

OR Not enough material to evaluate.

*NOTE: *Appropriate* is defined as addressing all aspects of a topic, for example, both advantages and disadvantages, or all characteristics in questions involving choices. Furthermore, *all* parts of the prompt should be touched on.

Language (Grammar, Vocabulary, Register, Mechanics)

9-10

a. Except for rare minor errors (esp. articles), the grammar is native-like.
b. There is an effective balance of simple and complex sentence patterns with coordination and subordination.
c. Excellent, near-native academic vocabulary and register. Few problems with word choice.

7-8

a. Minor errors in articles, verb agreement, word form, verb form (tense, aspect) and no incomplete sentences. Meaning is never obscured and there is a clear grasp of English sentence structure.
b. There is usually a good balance of simple and complex sentences both appropriately constructed.
c. Generally, there is appropriate use of academic vocabulary and register with some errors in word choice OR writing is fluent and native-like but lacks appropriate academic register and sophisticated vocabulary.

5-6

a. Errors in article use and verb agreement and several errors in verb form and/or word form. May be some incomplete sentences. Errors almost never obscure meaning.
b. Either too many simple sentences or complex ones that are too long to process.
c. May be frequent problems with word choice; vocabulary is inaccurate or imprecise. Register lacks proper levels of sophistication.

3-4

a. Several errors in all areas of grammar which often interfere with communication, although there is knowledge of basic sentence structure.
b. No variation in sentence structure.
c. Frequent errors in word choice (i.e., wrong word, not simply vague or informal word). Register is inappropriate for academic writing.

1-2

a. There are problems not only with verb formation, articles, and incomplete sentences, but sentence construction is so poor that sentences are often incomprehensible.
b. Sentences that are comprehensible are extremely simple constructions.
c. Vocabulary too simple to express meaning and/or severe errors in word choice.

OR Not enough material to evaluate.

Rhetorical Control

9-10

a. Introduction and conclusion effectively fulfill their separate purposes: The introduction effectively orients the reader to the topic and the conclusion not only reinforces the thesis but provides new insight.
b. Paragraphs are separate, yet cohesive, logical units. Sentences form a well-connected series of ideas of logical steps with clarity and efficiency.

7-8

a. The introduction presents the controlling idea, gives the reader the necessary background information, and orients the reader, although there may be some lack of originality in the presentation. The conclusion restates the controlling idea and provides a valid interpretation but may not provide new insight.
b. Paragraphs are usually logically developed and cohesive. Sentences are usually well-connected.

5-6

a. Introduction presents the controlling ideas but may do so mechanically or may not orient the reader to the topic effectively. The conclusion does not give the reader new insights or may contain some extraneous information.
b. Paragraphs are sometimes incompletely or illogically developed. Sentences may not be well-connected.

3-4

a. Introduction and conclusion do not restate the controlling idea. Introduction fails to orient the reader adequately, and the conclusion may not be tied to the rest of the essay.
b. Paragraphs are often incompletely or illogically developed and sentences are not well-connected.

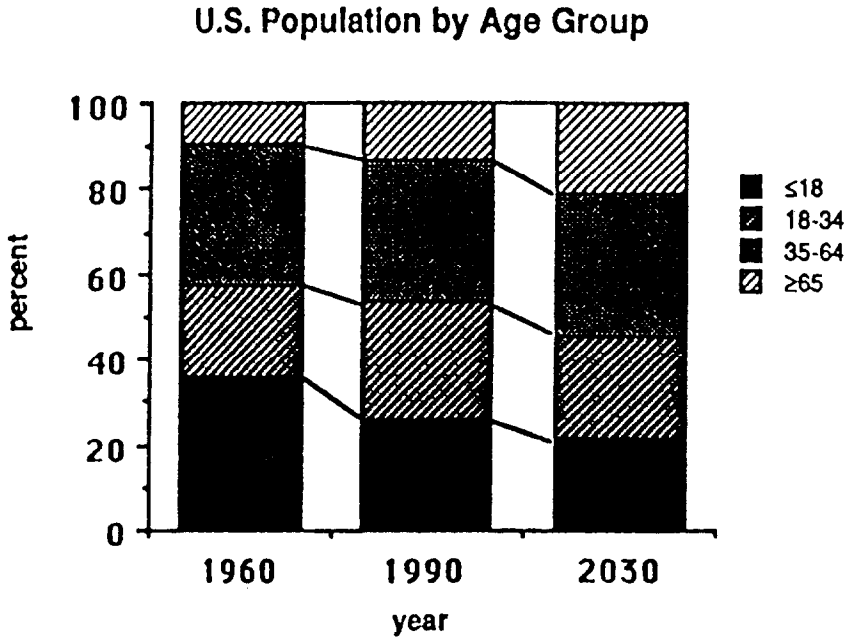
1-2

a. Introduction and conclusion are missing or unrelated to rest of the essay.
b. There is no attempt to divide the essay into conceptual paragraphs, or the paragraphs are unrelated and the progression of ideas is very difficult to follow.

OR Not enough material to evaluate.

APPENDIX B: COMPOSITION PROMPTS

Composition Topic 1



The above graph shows the percentage of people in different age groups in the United States population from 1960 to 2030. What does the graph tell you about changes in the population of the United States? What problems will people face as a result of these changes, and how can the best prepare for these problems?

Composition Topic 2

Imagine that you have been offered three jobs, and you must decide which offer to accept, based on the information below. Which job would you choose, and why? Discuss the advantages and disadvantages of your choice.

	Job A	Job B	Job C
Salary	$25,000	$40,000	$100,000
Hours/week	40	50	60
Vacation	6 weeks	3 weeks	1 week
Job Satisfaction	High	Medium	Medium