LLT Journal, e-ISSN 2579-9533, p-ISSN 1410-7201, Vol. 24, No. 1, October 2021

LLT Journal: A Journal on Language and Language Teaching
http://e-journal.usd.ac.id/index.php/LLT
Sanata Dharma University, Yogyakarta, Indonesia

THE EFFECTS OF NARRATIVE AND ARGUMENTATIVE MODES ON ASSESSING LEARNERS' WRITTEN PERFORMANCES BASED ON THE ANALYTIC RATING SCALE

Rania Zribi and Chokri Smaoui
Sfax University, Tunisia
raniazribi@ymail.com; smaoui2002@yahoo.com
correspondence: raniazribi@ymail.com
DOI: 10.24071/llt.v24i1.2986
received 19 November 2020; accepted 29 October 2021

Abstract
This study aims at investigating the effects of discourse modes on assessing EFL learners' written performances. A total of fifty raters judged sixty essays (30 narratives and 30 argumentative writing modes) written by third-year English students from the Faculty of Letters and Humanities. Raters not only scored the compositions but also justified their scores' assignments based on written explanations. Raters' rating behaviors were diagnosed based on a variety of quantitative and qualitative tools. Essay scores were analyzed based on the statistical model FACETS to measure raters' severity and internal consistency, task difficulty, and the scale functioning across writing modes. Qualitative data (gathered from interviews and report forms) were also analyzed in order to examine which aspects of writing were deemed more important than others across task types. The analysis revealed that the discourse mode was substantially an influential factor. The narrative task was more difficult than the argumentative one. Narrative essays were judged harsher than argumentative essays. Less consistent ratings could be detected from the narrative mode, compared to the argumentative one. Qualitative findings showed that the two writing modes were different in their qualitative judgments due to their different genre requirements and norms.

Keywords: discourse modes, scoring, rating scale, FACETS, scores' variability

Introduction
Academic writing is a crucial communicative skill in English as a first language (L1), second language (L2), and foreign language (EFL) teaching and learning instructions. It is a sophisticated "form of thinking" (Zinsser, 1988 p. vii), in which the writer has to perform different actions simultaneously, such as planning, organizing, writing, revising, editing, and publishing (Weigle, 2002 p.4) to produce coherent and accurate performance. The mastery of this complex skill is essential for university students, who are required to develop their writing abilities at this level through their cognitive process of not only constructing meaningful knowledge but also transmitting messages to their readers based on academic essays.

Apart from its complexity in language teaching, the writing skill also seems difficult to be tested by EFL teachers. In this respect, Bizzell (1987) focuses not only on the complex nature of the writing activity, but also on the extreme difficulty of assessing it, especially with the presence of human raters, whose great deal of subjectivity constituted one of the perennial problems related to direct writing assessments (p.583).

While reviewing the literature, it has been noted that the same learners' written performances were assessed subjectively by different raters despite the use of the same rating scale with its well-defined rating criteria, resulting in inconsistencies that would threaten not only reliability but also validity in the writing assessment context. Raters have different perceptions of what constitutes a good writing sample. What is appraised by one rater is downplayed by another. In justifying their scores' assignments, raters may overlook some mistakes while others may magnify them in measuring students' language skills. On that account, raters' potential scores' variability and divergent rating judgments can be due to various factors related mainly to raters, writing modes, rating scales, rating criteria…etc.

Out of the myriad influential sources of scores' variability, this paper attempts to focus on the task variable, because as Barkaoui (2008) claims, "task characteristics can also influence rater performance and reliability'' (p.12). It is thus proposed, in this work, to provide a deep analysis of the effect of task types on raters' quantitative and qualitative judgments.

Our main intention is to investigate and the way raters assess learners' written performances based on a well-defined analytic rating criteria. Our primary goal is to analyse the possible discrepancy in raters' judgments of narrative and argumentative writing modes in the analytic rating scale, by taking into consideration not only their severity and internal consistency rates but also the difficulty estimates of the two writing task types and the scale functionality. This research also will focus on the writing aspects that attracted the attention of raters in evaluating the same test takers' narrative and argumentative essays analytically.

The current study addresses the following research questions:

1. What are the effects, if any, of narrative and argumentative tasks on raters' severity and internal consistency based on the analytic rating scale?
2. To what extent do narrative and argumentative tasks vary in terms of their difficulty estimates?
3. Do writing modes influence the functionality of the analytic rating scale?
4. Do different task types affect the raters' scoring behaviors and the aspects of writing they attend to, based on the analytic rating scale?

Review of the literature

Discourse mode, a task characteristic that could potentially influence the assessment of EFL learners' writing performances, is of particular interest in the present study. A crucial issue pertaining to the evaluation of writing proficiency is scores variation among raters due to different variables, mainly the tasks variable. The latter should be controlled in testing writing skills to allow learners to generate their best performance and to ensure valid and reliable scores. In the realm of academic writing assessment and scores variation research, the effect of prompt types on raters' scores has been profusely investigated by researchers and

specialists in the field (Engelhard et al., 1992; Kegley, 1986; Kuhlemeir, et al., 1995; Quellmalz, et al., 1982; Sachse 1984). Several empirical researches have pointed to the conceivable impacts of different task requirements and modes on raters' scoring behaviors (Cumming et al. 2002; Weigle, 1999) and the reliability of their scores assignment (Tedick, 1990).

In this context, Stifler (2002) maintains that "modes of writing, or rhetorical modes are patterns of organization aimed at achieving a particular effect in the reader" (p.1). This idea was encapsulated by White (1982) in his claim that "we know that assigned mode of discourse affects test score distribution in important ways. We do not know how to develop writing tests that will be fair to students who are more skilled in the modes not usually tested" (p.17). To stress the inter-relation of both social and educational settings in language communication, Weigle (2002) focused on the effect of writing tasks and contextual factors on tests scores. She claims that "any assessment takes place in a given social and cultural context and may not be generalizable outside that context" (p.60). In this regard, Oxford (1996) states that "when language learners are asked to tell their histories, they inevitably address contextual, situational, cultural factors as part of the story of their learning" (p.582). Thus, Context has emerged as a vital theme in the educational system.

A possible source of scores variation examined in this paper after measuring EFL learners' essays was the discourse mode facet (narrative vs. argumentative tasks). The chief aim of previous research is to compare the raters' scores assignment to two or more writing modes to extract their points of similarities and differences in essays measurement. The research finding of Kegley's (1986) study can be used to illustrate the considerable effects of discourse mode on the assessment of the writing competence. She perceived differences between the mean score of a narrative sample and marks for descriptive, expository, and persuasive samples. The narrative essays received the highest marks while the persuasive essays received the lowest marks (p.147).

In one of the studies investigating the notable effect of discourse mode on writing scores assignment task, Engelhard et al. (1992) proclaim that "narrative writing tasks received the highest ratings, with descriptive writing tasks receiving the next highest rating, and expository writing tasks receiving the lowest ratings" (p.329). In examining raters' scores to two different discourse modes, the findings of Carrell's (1995) study were condensed to denote higher holistic scores assigned to the narrative essays than to the argumentative essays produced by the same writers (p.175).

In contrast, Quellmalz et al. (1980) obtained nearly opposite results in examining the relationship between two discourse modes and raters' scores to learners' compositions. They found that scores assigned to narrative essays were lower than those given to expository essays based on a five-point holistic rating rubric. Raters' variability can be due to their tendency to rate narrative mode of discourse in a stringent way or to the examinees' lack of knowledge or to their curricula requirements (p.13). Moreover, a strong correlation can be detected in the scores assigned to two essays produced in the same mode of discourse. However, that was not the case with the scores awarded to two essays in different discourse modes (p.13). In another study conducted to diagnose the effect of discourse modes on raters' testing or writing quality, Quellmalz et al. (1982) concluded that "levels of performance vary on tasks presenting different writing

purposes" (p.255). Hence, this divergence in scoring the two different discourse modes could be attributed to task requirements, as each task requires different writing skills, leading to construct-relevant variance, which causes aberrations in raters' score assignment tasks.

Method

Research design overview

This study adopted a cross-sectional design to gather sufficient data from the examinees and their raters at a single point in time. Its chief aim is to analyse and interpret English raters' evaluative behaviors and scores' assignments when testing EFL learners' writing responses to two different narrative and argumentative discourse modes. Hence, both teachers and students took part in this empirical research.

A comparative pattern was also incorporated in this study to extract the differences and similarities in the scores and judgments assigned by raters to EFL test takers' writing samples on two different writing modes based on the analytic rating scale. To advocate the efficiency of the comparative design in analysing the study outcomes in the language testing field, Collier (1993) argued that "comparison is a fundamental tool of analysis. It sharpens our power of description, and plays a central role in concept-formation by bringing into focus suggestive similarities and contrasts among cases" (p.105).

Participants

A total sample of thirty EFL learners voluntarily took part in this study. They are, a representative sample of a large population, enrolled in the third year English class level. These students were mostly females with a mean age of 22. They were under-graduate third-year English students, who have been specialized in the target foreign language for three years at the tertiary level and whose proficiency levels vary. All the test takers were non-native speakers of English and students in the English department at the xxx university. In addition, a panel of fifty writing teachers of English as a foreign language participated in this phase. They represented a mixed sample of male and female raters with an average age of 45 and belonged to different L1 backgrounds. Their first language is Arabic while English is their dominant work language in tertiary education in different Tunisian universities; they are specialized in teaching English as a Foreign language to EFL under-graduate learners.

At the time of my data collection process, all third year learners were attending their English classes and lessons. From the third-year class, I selected randomly female and male students to sit for two separate one-hour task-based writing performance tests on two different testing occasions to respond silently to two different discourse modes, viz. narrative and argumentative prompts. Thus, each test taker produced two writing samples, to come up finally with a total of sixty essays. In the first task, each test taker is required to write an essay in which he narrates the way he has helped his family to solve a family problem. On the other hand, in the argumentative task, candidates were asked to provide their arguments to convince the reader about the assets and drawbacks of using technology in our society.

The next step consisted in collecting the examinees' writing performances. To control the effects of such variables as handwriting, these samples were typed,

without changing or removing the original mistakes. As it was neither a part of the learning objectives of the writing course at the third year level nor mentioned in the analytic rating rubric that raters relied on, handwriting was not among the writing aspects to be tested in the data collection phase. The names of the students, who composed the two essays were also removed and replaced with just numbers in order to minimize potential bias rates. The sixty narrative and argumentative essays (thirty essays in each task) were sent to fifty raters to judge and score their quality based on the same rating instruments and procedures.

Procedures

Test takers were instructed to generate an essay after responding to each task requirements on two different testing occasions. First, each candidate produced an essay after responding to the narrative prompt, then, within a one-week period, he responded to the argumentative writing task by generating an argumentative sample. The time allowed for each essay production was one hour to enable students to understand the topic, analyze it and generate a coherent writing sample in an authentic testing context. The present study used two different essay prompts, which vary in terms of their characteristics, notably content, structure, and wording. They are designed to evaluate EFL examinees' abilities to perform coherent and well-structured academic writing samples.

Written production was prompted via computer and then sent to raters to judge students' writing proficiency in the two discourse modes based on the same analytic rating scale. In this respect, the ESL Composition Profile designed by Jacob et al. (1981), was applied in this study to test written productions. This analytic scale was originally constructed for large-assessment purposes to test multiple composition samples of English as a second language. It comprises five different criteria, namely content, organization, vocabulary, language use, and mechanics (See Appendix A). Since this study has embarked on examining raters' scoring patterns of EFL learners' writing performances in both narrative and argumentative discourse modes, a mixed-methods triangulation design of both quantitative and qualitative approaches was applied to gather and report data about the raters' decision-making while rating under-graduated learners' essays.

Quantitative data were extracted from different procedures. Analytic score report forms were employed to come across raters' scores assignment and their decision-making process after assessing EFL learners' writing compositions (See Appendix B). These analytic scores awarded by raters to the same test takers' narrative and argumentative essays were analysed based on two statistical programs: SPSS and FACETS (version 3.80.0). The latter permits researchers to add as many facets as they need, such as raters, rater groups, tasks, students, rating scale, rating criteria, and so on depending on the purpose of each study. In this vein, Schaefer (2008) highlights the prominent value of this model by stating that "it has shown great promise in the area of performance assessment and rating scale validation because it can analyze sources of variation in test scores besides item difficulty or person ability" (p.466). Prior to the analysis and interpretation of raters' judgments, the facets used in this study were coded. The following figure presents the relevant facets related to this study in the data collection phase.

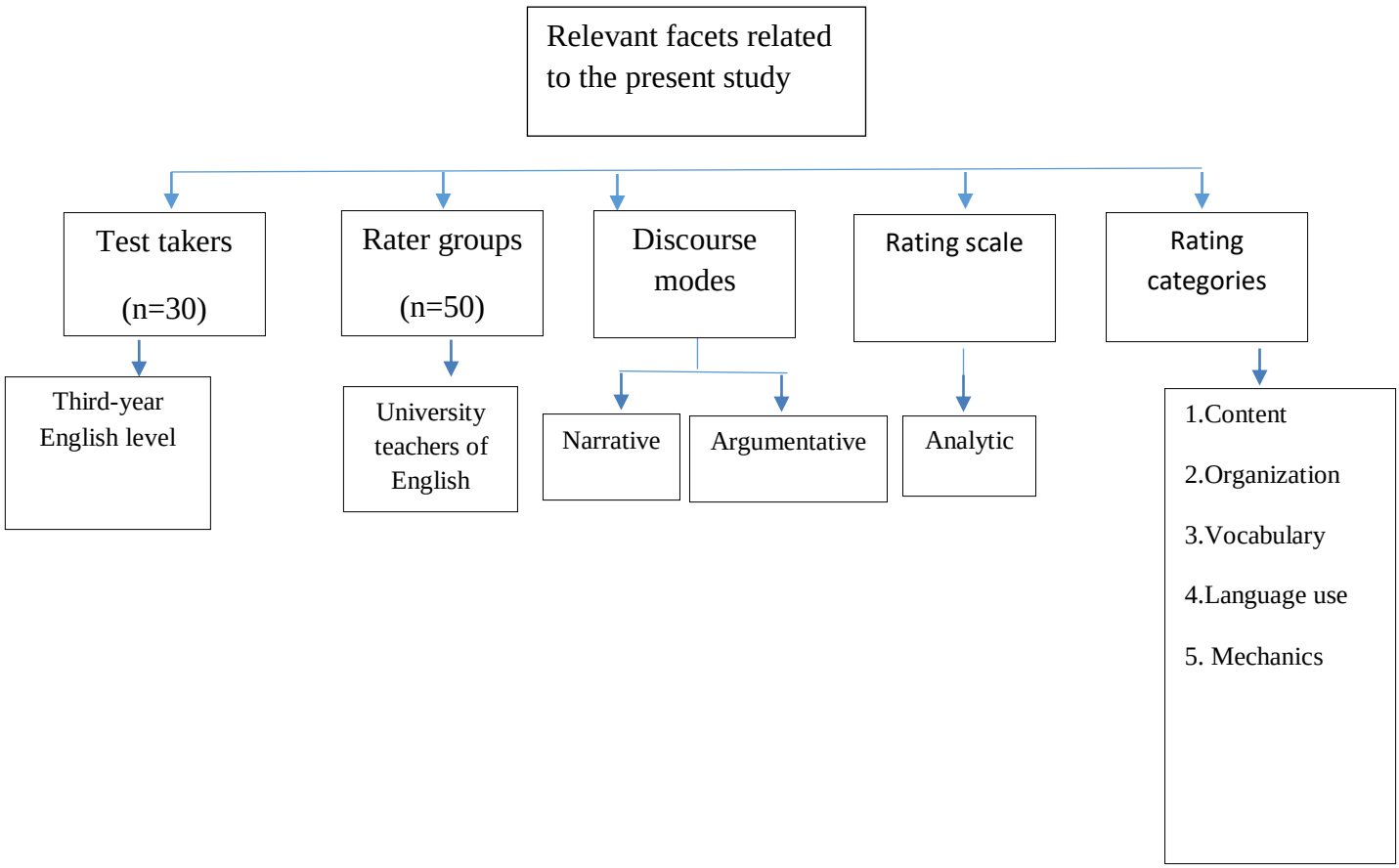

Figure 1: A schema of the relevant facets of assessment to this study

A profile questionnaire about raters' qualifications, and personal and professional data was also adopted (See Appendix C). These statistical procedures are used in order to examine the effects of evaluating different narrative and argumentative writing modes on raters' severity and internal consistency, task difficulty, and rating scale's performance. Furthermore, report writing forms were conducted to gather qualitative data about raters' judgments of EFL test takers' writing responses to two different narrative and argumentative prompts. Raters' reading and assessing strategies were thus elicited based on explaining their rating patterns during the evaluation process.

Findings and Discussion

Analyzing raters' quantitative judgments across tasks

Both FACETS and SPSS statistical outcomes across task types were reported in this section to analyse raters' scoring behaviors based on their analytic marks assigned to the same candidates' narrative and argumentative essays.

Rater severity

FACETS analysis revealed clear differences in raters' severity levels after assessing narrative and argumentative performance analytically. Measuring rater severity/ leniency levels on a logit scale, centred at 0, spanned 4.09 logits, from the most lenient rater located at -2.03 to the most severe rater located at 2.06 for

the narrative task and a 3.32 logit spread, with the most lenient rater at -2.30 to the harshest rater at 1.02 for the argumentative task. Comparing the range of raters' severities across task types, the table shows raters' variations.

It is clear from table 1 that raters differ in their severity estimates. They graded the narrative essays more severely than the argumentative essays (A span of 4.09 for the narrative task is smaller than the 3.32-logit spread for the argumentative task). This may be because of the judges' tendency to mark narrative essays more stringently (Quellmalz et al. 1982 p.13) or due to their perceptions of the writing task difficulty and their attempts to adjust scoring behaviors accordingly. These results were at variance with Engelhard et al'. (1991) study, which showed that "writing tasks that require more personal responses (direct and imagined experiences) tend to elicit essays that receive higher ratings than writing tasks that require impersonal or outside knowledge" (p.19).

To test the significance of these raters' different severity levels across tasks, FACETS output generated three indices, namely the separation statistics, chi-square with its p-value and the reliability of separation.

Table 1: Summary of Rater Measurement Report by Discourse Modes (based on the analytic scoring procedure)

	Narrative Mode	Argumentative Mode
Rater Severity		
M (Model SE)	.13	.13
SD (Model SE)	.00	.01
Min	-2.03	-2.30
Max	2.06	1.02
Infit		
M	1.0	1.0
SD	.25	.31
Outfit		
M	1.01	1.0
SD	.25	.33
Separation Statistics		
Separation Ratio (G)	8.75	7.85
Separation Index (H)	12.00	10.80
Reliability of Separation	.99	.97
Fixed chi-square statistics	3519.0	2868.8
df.	49	49
Significance	.00	.00
Inter-rater agreement opportunities	90000	90000
Exact agreement %	37.8% 34044	38.4% 34516
Expected agreement %	35.9% 32305.2	36.8% 33129.3

Analyzing raters' analytic scoring decisions across the two tasks was based on FACETS outcomes as illustrated in the table above. The item separation ratios (G) were 8.75 for the narrative task and 7.85 for the argumentative task, indicating that the variance among scorers was approximately nine times higher than the error of estimates, especially for the narrative essays, thus suggesting that graders were not equally severe. The prompt separation index (H) was 12.00 for the

narrative essays and 10.80 for the argumentative essays, indicating that raters can be divided into about twelve severity levels in the narrative task and eleven levels in the argumentative task.

Separation statistics, separating raters into more distinct severity levels for the narrative prompt (twelve levels) than for the argumentative prompt (eleven strata of severity), were fairly reliable due to high reliability of separation indices of .99 for the narrative task and .97 for the argumentative task. Thus, raters showed significant notable differences in the levels of severity they exercised for the two tasks, with a fixed chi-square value of 2868.8 for the argumentative task and 3519 for the narrative task (degree of freedom = 49) and a significant p-value at .00 ($p < .005$). The null hypothesis that all scorers were equally harsh in their scores' assignment to the candidates' narrative and argumentative writings must be rejected.

Measuring inter-rater reliability rates in assigning marks to the students' narrative and argumentative essays was based on inter-rater agreement statistics. As table 1 demonstrated, out of 90000 possible opportunities for agreement, the numbers of exact agreements between raters were 34044 (37.8%) for the narrative essays and 34516 (38.4%) for the argumentative essays, while the expected ones were 32305.2 (35.9%) and 33129.3 (36.8%) for the narrative and argumentative writings respectively. The observed exact agreements (37.8% and 38.4%) were higher than the expected percentages (35.9%) and 36.8%). This explained the fact that raters did not judge their test takers' performance in an independent way. Task types remain an influential factor in assessing learners' writing skills in different writing modes.

Rater internal consistency

A more detailed analysis of raters' internal consistency in judging examinees' narrative and argumentative performance is based on fit statistics in the rater facet. A preferable infit mean-square value of 1.00 was perceived in the two prompts, suggesting intra-rater agreement between raters in assessing both tasks based on the analytic rating scale. They not only employed the analytic rating scale consistently but also maintained their severity levels across the two tasks. Their internal consistency in measuring learners' writing modes could be explained by the fact that their scores assignment fitted perfectly the Rasch model predictions.

Little variation however can be detected in the outfit mean-square values of 1.01 for the narrative task and 1.0 for the argumentative task. To further investigate raters' internal consistency across discourse modes, the same three-class fit pattern of overfit, acceptable fit, and misfit was analysed to differentiate between raters in terms of their intra-rater agreement rates in measuring examinees' samples. The following table exhibits the frequencies of raters' consistency in the analytic assessment of the same set of narrative and argumentative essays.

Table 2: Frequencies of Rater Fit Statistics across Discourse Modes (based on the analytic scoring procedure)

	Narrative Mode		Argumentative Mode	
Fit Range	Infit MS	Outfit MS	Infit MS	Outfit MS
Overfit: Fit < 0.70	3 (6%)	5 (10%)	8 (16%)	9 (18%)
Acceptable Fit: 0.70 < Fit <1.30	36 (72%)	32 (64%)	40 (80%)	39 (78%)
Misfit : > 1.30	11 (22%)	13 (26%)	2 (4%)	2 (4%)

Out of the fifty raters, thirty-six (72%) exhibited acceptable infit estimates in measuring the test takers' narrative performance, while forty (80%) displayed acceptable consistency rates in testing the argumentative writings based on FACETS output. Raters thus showed slightly higher intra-rater agreements in rating the argumentative tasks as compared to the narrative tasks. There were more overfitting raters (n= 8, 16%) in judging the argumentative essays compared to the narratives (n= 3 representing 6%), indicating little variation between scores' assignment process to the two tasks and the FACETS expected scores. More misfitting raters (n= 11) appeared in scoring the narrative essays (22%) than in marking the argumentative essays (n= 2 representing 4%), suggesting much variability in the marks awarded to the narrative samples. Misfitting raters, whose different rating behaviors did not fit the model, threatened scores validity in the testing field.

A small number of misfitting raters across tasks appeared in the analytic ratings (4% for the argumentative task and 22% for the narrative task). Based on these outcomes, raters were more consistent than the model predicted in scoring the argumentative essays, compared to the narrative essays. This can be attributed to the open-ended personal nature of narrative essays, which are difficult for raters to judge consistently, leading to such unwanted scores' variations and inconsistent rating behaviors. The null hypothesis stating that raters across two distinct discourse modes showed the same severity and internal consistency levels in scoring the same test takers' writings based on the analytic rating scale must be rejected.

Prompts difficulty

After analysing the rater facet in terms of severity measures and internal consistency estimates, it is crucial to focus on the task facet, as one of the variables in the current study by taking into account both the difficulty estimate parameter and fit statistical indices for each task. The former was applied to measure the difficulty levels of the two tasks, while the latter was used to test the consistency of measuring these tasks difficulty rates. Task average difficulty is set at 0 logit by convention. Table 3 illustrates prompt difficulty measures for both discourse modes resulting from the analytic ratings.

Table 3: Prompt difficulty estimates (n=2) (based on the analytic scoring procedure)

	Observed Average	Fair Measure Average	Measure	Model SE	Infit MS	Outfit MS
Narrative	2.56	2.58	.14	.02	1.04	1.05
Argumentative	2.68	2.71	-.14	.02	.96	.96
Mean (n= 2)	2.62	2.65	.00	.02	1.00	1.00
SD	.09	.09	.20	.00	.05	.06

RMSE .02 Adj (True) S.D. 20 Separation 11.36 Reliability .99
Fixed (all same) chi-square: 130.0 d.f.: 1 significance (probability): .00

As can be drawn from table 3, the fair measure average for the narrative task (2.58) was less than the fair measure average for the argumentative (2.71) on the four-point analytic scale, which demonstrated that the narrative task was more difficult, as compared to the argumentative. To measure the underlying difficulty of the two prompts, a logit difficulty of .14 (SE = .02) for the narrative task was higher than the difficulty span of the argumentative task (-.14 with SE = .02), which indicated that it was more difficult to get a high score on the narrative task than on the argumentative one in grading essays based on the analytic rating rubric. The narrative task appears to be more difficult, compared to the argumentative task in the analytic rating scale. This can be explained by the fact that test takers are expected to perform their stories by narrating past experiences with special attention to correct language and rhetorical aspects of language to form coherent and fluent narrative flow. This personal open aspect of narrative flow reflects its difficulty for students and raters alike.

To test the significance of the different levels of difficulty between the narrative and argumentative prompts, the fixed chi-square test with its p-value were underlined. The fixed chi-square yielded a value of 130.0 with 1 degree of freedom and a significant p-value (= .00), which rejects the null hypothesis that the two tasks are equal in difficulty based on the analytic scorings.

Analytic scale functioning

From FACETS output, test takers' ability measures can be extracted to examine the functionality of the analytic rating scale with its five categories across discourse modes. The adequacy of the analytic rating scale with its five rating categories can be measured based on the threshold (step) calibration statistics generated from FACETS. For instance, concerning the content criterion, a test taker whose ability estimates was -2.11 for the narrative mode and -2.36 for the argumentative mode had a probability of 50% to be scored as either 2 or 3. Testing the scale functioning is also based on the way the category thresholds were ordered. According to table 4, the content category thresholds for example were ordered in an ascending order from -2.11 to 2.13 for the narrative mode and from -2.36 to 2.72 for the argumentative mode in the four score levels. The thresholds measures in the five rating categories increased monotonically as the score levels advanced. The distance between the thresholds was also adequate in the five rating categories, as they advanced by at least 1.4 logits from one score level to another, but did not exceed 5 logits. This indicates that the analytic rating categories were not only ordered but also functioned as expected by the model as

the thresholds advanced monotonically with the analytic scale levels. This stressed the functioning of the analytic scale.

Table 4: Threshold Estimates for the Analytic Scoring Categories across Tasktypes

The Narrative Mode					
Score Levels	Content	Org	Vocab	Lg Use	Mech
1	None	None	None	None	None
2	-2.11	-1.79	-2.37	-2.29	-1.91
3	-.02	-.38	-.11	.09	-.09
4	2.13	2.17	2.48	2.21	2.00
The Argumentative Mode					
Score Levels	Content	Org	Vocab	Lg Use	Mech
1	None	None	None	None	None
2	-2.36	-1.86	-2.52	-2.32	-1.89
3	-.36	-.51	-.28	-.15	-.12
4	2.72	2.38	2.80	2.47	2.01

Aspects of writing attended to across task types

The table below presents the percentages of referring to each rating category in the analytic report forms after assessing learners' narrative and argumentative essays. It shows that overall, raters explained their scores based not only on each rating category of the analytic scale but also on other writing aspects not mentioned in the scale. The highest percentages pertained to both content (28.67% for narrative essays and 27.82% for argumentative essays) and organization (about 22% for narrative essays and 28% for argumentative essays) across task types, while the lowest concerned mechanics (about 6% for narrative essays and 5.80% for argumentative essays) and vocabulary (about 12% for narrative essays and 10% for argumentative essays) criteria.

Other writing aspects were more reported in the narrative mode (15.32%) than in the argumentative mode (12.40%). Additionally, raters reached approximately the same percentage (16% and 15.64%) in using the language use aspect across the two writing modes. Based on the frequency of raters' comments across the five rating criteria to the narrative and argumentative essays, we can deduce that both writing modes reported the five rating criteria together with other writing aspects based on the same order of importance but with slight differences in terms of their percentages. Hence, content, organization, language use, other aspects, vocabulary and mechanics were mentioned in both tasks with some differences in frequency ranges.

Table 5: Frequencies for Aspects of Writing in Report Forms across Task Types

	Content	Org	Vocab	Lg Use	Mechanics	Other Asp
Narrative Mode	28.67	21.91	11.91	16	6.17	15.32
Argumentative Mode	27.82	27.93	10.41	15.64	5.81	12.40

To scrutinize whether the differences in percentages of decision-making strategies and aspects of writing reported in the analytic report forms across task types were significant, I conducted a Wilcoxon Signed-Ranks test for each rating category. The table below portrays the statistical significance of the difference in using the six rating criteria across narrative and argumentative tasks based on the non-parametric Wilcoxon signed-ranks test. The Wilcoxon test indicated that only the difference for organization was significant as p-value is below the threshold level of 0.05 (p = .000). The p-values of each of the other five rating categories, being .748, .169, .922, 724, and .009 are all above the threshold of 0.05. Therefore, it was clear that the use of the five scoring criteria was not significantly different across writing prompts. Different task types did not significantly influence the raters' score assignment and decision-making tasks.

Table 6: Test Statistics across narrative and argumentative tasks

Test Statistics[a]	Content	Org	Vocab	Lg Use	Mechanics	Other Aspects
Mann-Whitney U	457440,000	402240,000	448800,000	459840,000	458400,000	436320,000
Wilcoxon W	918720,000	863520,000	910080,000	921120,000	919680,000	897600,000
Z	-,322	-5,568	-1,376	-,098	-,353	-2,611
Asymp. Sig. (2-tailed)	,748	,000	,169	,922	,724	,009

a. Grouping Variable: Task Type

The above quantitative data analysis explicates the effect of narrative and argumentative writing modes on raters' scoring patterns based on the analytic rating rubric. The statistical FACETS outcomes, highlighting the impact of discourse modes on the scores' assignment task, showed raters' different decision making processes across task types, which will be further examined in the qualitative part. This section clarifies the aspects of writing that raters attended to across the narrative and argumentative tasks, by analyzing report form explanations, associated with the writing modes variable.

Analyzing raters' qualitative judgments across tasks

During the assessment process, the fifty scorers were required to explain their assigned marks to the sixty students' performance by writing their remarks in the analytic report forms. Raters' comments were then classified across narrative and argumentative writing modes. This classification helps us to compare raters' explanations across tasks, by examining which aspects of language may attract the raters' attention in evaluating narrative and argumentative writings. This qualitative analysis thus investigates the nature of the raters' scoring behaviors across tasks. A scrutiny of raters' feedback suggested that raters' score explanations to narrative performances outnumbered argumentative essays based on the four-point analytic rating scale with its five different rating criteria. This

can be due to the difficulty of the narrative task and to the test takers' inability to generate coherent and accurate narrative performances.

In terms of content, judges were aware of genre stipulations and topic requirements across tasks. Although clear explanations, examples and details enhance ideas elaboration and arguments development, different task-specific functional elements affected raters' judgments of both tasks. Judges referred to the problem, the solution and the consequence in measuring narrative essays, by commenting on chronological story events, its plot and climax. Raters, however, pointed to issue, evidence, claim, counter argument and support in argumentative essays, by focusing on the development of the thesis and the anti-thesis and parallelism between both parts of argumentation. In terms of focus, raters directed their attention to the use of quotations in the narrative mode and references in the argumentative mode. It seems that content affects not only coherence but also organization in narrative prompts.

In terms of organization, raters provided comments on the conventional five-paragraph essay structure format with respect to both task types. Despite the importance of paragraphing in the two modes, argumentative essays were evaluated based mainly on the specific components of an introduction. Raters valued the presence of motivator, background information, thesis statement and blueprint in producing a well-structured argumentative introduction. They even focused on topic sentences in each controlling paragraph. They extended their comments to focus on recommendations in the conclusion. For the narrative tasks, however, raters pointed to the effect of fragmented ideas on developing the narrative flow of events, leading to unbalanced narrative performance, which in its turn resulted in an incoherent story. Both connection of ideas and transition between the different essay parts are pertinent in the narrative and argumentative productions. These divergences can be attributed to the different narrative and argumentative structures and organizational components and essays' format.

Some noticeable vocabulary differences between task types appeared in raters' analytic scores explanations. While raters commented on vocabulary sophistication and variation across the two tasks, they made reference to the test takers' word and form choices, especially verbs, adjectives, and adverbs to express their ideas clearly in narrative essays. Such comments were related to word placement, comparative and plural forms to formulate well-structured argumentative essays. Narrative tasks are associated with the informal register while argumentative tasks are related to the formal register.

Differences in the way language use is valued by raters across tasks are also examined. The use of accurate language was commented on in the two tasks. Sentence constructions were a prevailing aspect in foreign language productions due to their complexity and forms variations. In response to narrative tasks, scorers referred to clear structures and expressions, similes, comparative forms, and parallel sentences, whereas in argumentative essays, they focused on word order and placement, and accurate well-formed sentences that reflect clear arguments. Conjunctions, long sentences, run-ons, fragments, prepositions, articles, pronouns and active voice can be found in both genres. A colloquial informal style was attributed to narrative essays, which was not the case with argumentative essays. More lexical and language mistakes appeared in the narrative task, compared to the argumentative task. Narrative essays are normally written in the simple past as test takers narrate their past experiences or

fictionalize stories, while argumentative tasks should be performed in the simple present tense. This seems to be due to the specific-task based assessment task. Raters may have some stylistic, syntactic, and lexical preferences associated with each writing mode.

Raters also appeared to focus on mechanics in measuring examinees' narrative and argumentative performances. Based on the analytic report forms, the majority of raters' comments were negative, indicating serious problems in capitalization, punctuation, linkers and spelling in both tasks. These serious mistakes led to awkward structures, unclear sentence boundaries and incoherent text, which hindered the quality of the essay.

After dealing with all the rating categories within the analytic rating scale, we noticed more comments related to writing aspects other than those mentioned in the rubric; content, organization, vocabulary, language use and mechanics. This can be explained by the fact that raters did not stick to the analytic rating scale during their measurement process of the sixty narrative and argumentative essays. This can be traced back to the possible specific task features that may attract raters' attention. In this respect, raters may focus on some narrative or argumentative aspects over others.

In assessing examinees' narrative writing abilities, raters started by expressing their overall impressions of the whole piece of writing, by taking into account the originality of students' narrative academic essays. In judging argumentative essays however, raters directed their attention to overall structure and relevance rather than to its original arguments. Narrative writings received more score explanations in grammatical and stylistic writing aspects, compared to argumentative essays. Raters commented on test takers' use of contractions, modals, pronouns and verbs in their narrative performance. They even pointed to tense problems as some test takers used the simple present in narrating events.

Raters were even more interested in perceiving stylistic features in both tasks. They referred to oral-like and colloquial narrative style of writing, which was not the case for formal argumentative essays. Argumentative style could be hampered by plagiarism. Judges also pointed to the effects of language interference and translation on both narrative and argumentative writings. Thus, differences pertaining to the above mentioned categories would appear to call for separate rating scales, related to task-specific feature and raters' focus, which had been the most frequently investigated in foreign writing assessment. Raters' cognitive processes and rating behaviors in narrative and argumentative tasks were different depending on their scoring approaches and their treatment of the rating criteria and other relevant writing aspects. These qualitative outcomes were in line with Quellmalz et al. (1982), who stated that "the different subskills included in the scoring rubric seem definitely to interact with discourse mode and, at the same time, to varying degrees are independent sources of variation in student writing performance" (p.20).

Conclusion

This mixed-method research examined the effects of discourse modes (narrative and argumentative essays) on assessing EFL learners' writing skills. The narrative essays were rated significantly harsher than the argumentative essays. In terms of raters' internal consistency across task types, the narrative writings led to a higher proportion of judges with fit statistics within the misfit

range, whereas the argumentative writings resulted in a higher proportion of scorers with overfit. Argumentative ratings were thus more consistent than narratives. The statistical program FACETS revealed that the narrative prompt was more difficult than the argumentative prompt. The qualitative outcomes were also complementary. Raters attended to different writing aspects across task types. Hence, the qualitative assessments were divergent across task types, indicating that raters held different conceptions of what constitutes a good writing.

The major findings of the current study can direct our attention to different implications. These immutable differences across tasks may require the use of task-based rubrics, related to the specificities and characteristics of each writing genre, one for the narrative and one for the argumentative. It is also important to highlight the necessity for test developers to use different task types in any EFL writing assessment context in order to gather various writing samples, which represent each test takers' writing ability and thus ensuring valid results "… by giving a broader basis for making generalizations about a student's writing ability" (Read 1991 p.87). The analytic scale is also recommended as it might be useful for diagnostic and placement aims in high-stake writing assessments to ensure valid and reliable outcomes. As the present study shows, raters varied in their severity and consistency levels in their scores' assignment across tasks. One strategy to improve this limitation is to apply the multi-faceted Rasch measurement model (FACETS) to adjust raters' marks and enhance raters' inter and intra-rater reliability in judging different task types. This statistical program, as Prieto and Nieto (2014) claim, allows the "analysis of the actions of different raters on different tasks, and enables us to determine, in part, whether the scoring categories appearing on rubrics must be adjusted or changed in order to obtain more consistent or valid scores" (p.386).

References

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes*. Unpublished doctoral dissertation. University of Toronto, Canada.

Bizzell, P. (1987). Review: What can we know, what must we do, what may we hope: writing assessment. *College English, 49*(5), 575-584.

Carrell, P., L. (1995). The effect of writers' personalities and raters' personalities on the holistic evaluation of writing. *Assessing Writing*, *2*(2), 153-l90.

Collier, D. (1993). The comparative method. In A. W. Finifter (Ed.). *Political Science: The State of the Discipline 2*. American Political Science Association.

Cumming, A., Kantor, R. & Powers, D., E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, *86*(1), 67-96.

Engelhard Jr, G., Gordon, B., & Gabrielson, S. (1991). *Writing Tasks and the quality of student writing: Evidence from a statewide assessment of writing*. Paper presented at the Annual Meeting of the American Educational Research Association.

Engelhard, G., Gordon, B., & Gabrielson, S. (1992). The influences of mode of discourse, experiential demand, and gender on the quality of student writing. *Research in the Teaching of English*, 26, 315-336.

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

Jacob, H., Zinkgraf, S., Wormuth, D., Hartfiel, V. F. & Hughey, J. (1981). *Testing EFL cmposition: A practical approach.* rowley. MassNewbury House.

Kegley, P., H. (1986). The effect of mode discourse on student writing performance: Implications for policy. *Educational Evaluation and Policy Analysis*, *8*(2), 147-154.

Kuhlemeier, H., van den Bergh, H., & Wijnstra, J. (1995). *Multilevel factor analysis applied to national assessment data.* Paper presented at the Annual meeting of the American Educational Research Association, San Francisco.

Oxford, R., (1996). When emotion meets (meta) cognition in language learning histories. *The Teaching of Culture and Language in the Second Language Classroom*, 581-594.

Quellmalz, E., Capell, F, & Chou, C., P. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement*, *19*, 241-258.

Quellmalz, E., Capell, F., J., & Chou, C., P. (1980). *Defining writing: Effects of discourse and response mode*. CSE Report No.132. University of California.

Sachse, P. (1984). Writing assessment in Texas: Practices and problems. *Educational Measurement*: *Issues and Practice*, *3*, 21-23.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*, 465-493.

Stifler, B. (2002). *Rhetorical modes*.

Tedick, D., J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes*, *9*, 123-143.

Veal, L. R. & Tillman, M. (1971). Mode of discourse variation in evaluation of children's writing. *Research in the Teaching of English*, *5*(1), 37-45.

Weigle, S., C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, *6*(2), pp. 145-178.

Weigle, S., C. (2002). *Assessing writing*. Cambridge University Press.

Weir, C., J. (1990). *Communicative language testing*. UK: Prentice Hall.

White, E., M. (1982). Some issues in the testing of writing. *Notes from the National Testing. Network in Writing*. New York: Instructional Resource Center of CUNY.

Yunick, S. (1997). Genres, registers and sociolinguistics. *World Englishes*, *6*(3), pp. 321-336.

Zinsser, W., K. (1988). *Writing to learn*. New York: Harper & Row.