Determining which type of rating scale should be used is a major assessment decision for educators (Hamp-Lyons, 1991; Weigle, 2002). This decision is critical because, as Weigle (2002) stated, "the score is ultimately what will be used in making decisions and inferences about writers" (p. 108). Even though assessment purpose and use, as well as other components of the writing test like task type, narrow down one's choices, in most assessment contexts educators still have

to choose among two or more rating scales. Nevertheless, there has been little empirical research on how different scale types affect second language (L2) essay rating processes and scores, and how raters perceive them (Hamp-Lyons & Kroll, 1997; Weigle, 2002).

Rating scales can be classified in terms of approach and scoring method. Cooper (1977) distinguished between holistic approaches, which include "any procedure which stops short of enumerating linguistic, rhetorical, or informational features of a piece of writing," and analytic approaches whereby the reader is "required to count or tally incidents of the features" (p. 4). Holistic approaches are further divided into multiple-trait, holistic, and primary-trait scoring methods. Multiple-trait scoring involves assigning multiple subscores to individual traits or dimensions (e.g., language, organization) and then summing those subscores to arrive at an overall score. In holistic scoring, the rater considers individual elements of performance but chooses one score to reflect the overall performance. For primary-trait scoring a single score is assigned to an essay according to the degree to which the writer has addressed the specific requirements of the task (Cooper, 1977; Goulden, 1992, 1994; Hamp-Lyons, 1991; Weigle, 2002). As Goulden (1992) explained, these differences in scoring methods reflect different assumptions about the relationships between the parts and whole of the performance or product being assessed. Thus, multiple-trait scoring assumes that "the sum of the subscores for the parts is exactly equal to a valid score for the whole and, by evaluating the parts, the rater has evaluated the whole." Holistic scoring, by contrast, assumes that the product or performance being assessed is "a whole entity" that should be judged as such since "the whole is not equal to the sum of the parts." Rather, "the whole is equal to the parts *and* their relationships" (p. 265).

Based on these differences in scoring methods and assumptions Goulden (1994) proposed two hypotheses concerning the effects of rating scales on the reliability and validity of performance scores. Goulden contended that "inter- and intra-rater reliabilities should be facilitated by highly structured guides that reduce personal choice." Multiple-trait methods should limit the traits to just those on the instrument and control the level of importance the rater gives to a trait. Holistic scoring, in contrast, allows raters to "include traits not listed and use personal judgment to determine how important a specific trait is to the overall score" (p. 74). In terms of validity, because multiple-trait scoring limits the traits that a rater may consider to those in the instrument, it may force raters to ignore important criteria or aspects that may affect the overall quality of the performance. Holistic scoring, in contrast, could include all relevant aspects since "raters can adjust the overall score to accommodate any aspects" of the performance/product, which will result in "an idiosyncratic set of supplemental traits different from those written in the basic guide" and subsequently lower inter-rater reliability (p. 74).

The literature provides guidelines for developing and validating rating schemes (e.g., Brown & Bailey, 1984; Hamp-Lyons & Henning, 1991) as well as arguments for and against different rating schemes (e.g., Hamp-Lyons, 1991, 1995; Perkins, 1983). For example, Perkins (1983) argued that while it is weak in reliability, holistic scoring has high validity when overall attained writing proficiency is the construct to be assessed. Multiple-trait scoring, in contrast, enhances reliability but lacks in practicality and is of questionable validity because it isolates text features from context. Hamp-Lyons (1991, 1995), on the other hand, contended that while holistic scoring is appropriate for scoring first-language (L1) essays, multiple-trait scoring has higher validity and reliability when rating L2 essays because different learners have different levels of proficiency in different aspects of L2 writing. Furthermore, multiple-trait scoring provides teachers and students with more feedback on students' performance.

Few studies have empirically examined the impact of holistic and multiple-trait scoring methods on L1 and L2 writing scores and rater reliability. Comparing the holistic and multiple-trait

scores of college students and expert writers, Freedman (1979) found that the expert writers were distinguished from the college writers on the multiple-trait scale but not the holistic scale. And while the college writers received roughly the same scores regardless of the scale used, the expert writers consistently scored higher on the multiple-trait scale. Swartz et al. (1999) used Generalizability (G-) theory to investigate the reliability of writing scores derived from holistic and multiple-trait scoring of L1 narrative essays. They found that increasing the number of raters increased score reliability and that some rating scales (e.g., style, legibility) required more ratings to achieve acceptable reliabilities. Schoonen (2005) used G-theory to examine the impact of holistic and analytic scoring on the scores assigned to L1 essays in terms of language use or content and organization. He found that the writing tasks contributed more to score variance than the raters, but the generalizability of writing scores and the effects of raters and topics were dependent on the way the essays were scored (holistic or analytic) and the trait that was scored (language use or content and organization).

In L2 writing assessment, Carr (2000) examined how multiple-trait and holistic rating scales affect scores in an English as a Second Language (ESL) test that includes a writing component. The results of factor and regression analyses indicated that altering the rating scale changed the interpretation of the writing test, resulting in total test scores that were not comparable as the factor structure of the test itself changed. For the writing component Carr concluded that "the difference between [multiple-trait] and holistic scales is principally one of focus: holistic scores provide an assessment of a single construct, whereas composite scores from [a multiple-trait] rating scale conflate the information from several constructs" (p. 228). In a study investigating the effects of multiple-trait and holistic scoring on student placement in an English as a Foreign Language (EFL) program, Bacha (2001) found high correlations between the two sets of scores as well as high inter- and intra-rater reliabilities for both methods. Multiple-trait scoring, however, provided more information on students' performance in the different components of the writing skill. Other studies that considered the impact of different rating scales on essay scores and rater reliability found less consistency in the relationship between the holistic and multiple-trait scores assigned by inexperienced readers than between those assigned by experienced raters (Sweedler-Brown, 1985). Furthermore, the reliability of rating scales could vary considerably depending on the writing task involved (Schoonen, 2005; Schoonen, Vergeer, & Eiting, 1997).

I am not aware of any qualitative study that has examined the effects of different rating scales on essay rating processes. Most studies have investigated the decision-making behaviors and essay aspects raters attend to when marking essays with no specific rating guidelines (Cumming, Kantor, and Powers, 2002), or when using holistic (Milanovic, Saville, & Shuhong, 1996) or multiple-trait rating scales (Cumming, 1990; Lumley, 2002). Lumley (2002, 2005) may be an exception in that, although he did not specifically compare different scoring methods, his findings raise several relevant questions concerning the role of the scoring method in essay rating processes. Examining the rating processes of four experienced ESL essay raters, Lumley (2002) found that the raters faced problems reconciling their impression of the text, the specific features of the text, and the wordings of the rating scale. Second, although they seem to have understood the contents of the scale similarly in general terms, the raters might have applied them in different ways and emphasized different components of the scale descriptors. Third, the raters seemed to form their judgments independently of the scale wordings but "somehow managed to refer to the scale content" to articulate and justify their scoring decisions (p. 263).

The few empirical studies that specifically examined the impact of different types of rating scales on essay rating used quantitative methods to analyze essay scores. They did not consider whether and how the content and organization of rating scales impact essay-rating processes.

Such information is essential for designing, selecting, and improving rating scales and rater training as well as for the validation of L2 writing assessments. The present study combines both quantitative and qualitative methods to better understand the effects of different rating scales on L2 essay scores, rating processes, and raters' attitudes.

1. Background to the study

The present study was motivated by a desire to improve EFL writing assessment practices at a university in Tunisia. Currently, EFL writing evaluation at this university is done impressionistically, with raters assigning each essay a single score for its writing quality on a 20-point scale. Unlike the scoring methods described above, impression scoring includes no written scoring criteria or instructions. Instead, raters rely on their knowledge of the content of the EFL writing course they are teaching, their experience with EFL essay rating, and/or comparing essays to each other when assessing their students' EFL writing (i.e., a combination of criterion- and norm-referenced approaches). Usually, two raters independently rate each essay and their scores are then averaged to arrive at a final score. If the two raters assign "very different" scores to an essay, a third rater is called upon to rate it and the closest two scores are averaged.

Such an approach to essay scoring has at least two major disadvantages. First, because the scoring criteria are unique to each individual rater, large discrepancies in essay scores often occur. For example, one of the essays included in this study received the following scores from five EFL writing teachers: 6.5, 7.5, 8.5, 10, and 11. This variability reflects the idiosyncrasies of the individual raters and undermines the validity of the writing test as a measure of *students'* EFL writing abilities. While such variability in essay scores is not uncommon in other assessment contexts, the problem in this context is that there are no explicit criteria against which score discrepancies can be resolved. Second, teachers, particularly those new to the local context, may find it difficult to infer and implement these assessment criteria consistently. In fact, even experienced teachers might be employing different criteria and/or weighting essay aspects differently within and across rating sessions.

The study thus emerged out of a sense that introducing a rating scale with explicit and standard guidelines for evaluating students' EFL writing will improve the reliability and validity of this assessment. A key decision, however, was to determine which method should be used among holistic and multiple-trait scoring. In order to address this issue, the following questions were formulated:

1. What are the effects of holistic and multiple-trait rating scales on the dependability of the essay scores EFL teachers assign?
2. What are the effects of holistic and multiple-trait rating scales on EFL teachers' decision-making behaviors and the essay features they attend to?
3. How do EFL teachers perceive holistic and multiple-trait rating scales?

2. Method

2.1. Participants

The study included 32 essays on two tasks written by 16 volunteer intermediate EFL university students under exam-like conditions. The essays were rated by the four EFL writing teachers at the same university. Table 1 describes the four teachers, each of whom I refer to by a pseudonym.

Table 1
Profile of four EFL writing teachers

	Hatem	Nader	Paul	Fadwa
Gender	M	M	M	F
First language	Arabic	Arabic	English	Arabic
Degree	MA, Linguistics	MA, Cultural Studies	BA, Religious Studies	MA, English Literature
EFL teaching experience	6 years	14 years	6 years (ESL)	21 years
Rating experience	6 years	5 years	0.5 year	5 years

Three teachers, Hatem, Nader, and Fadwa, have a long EFL teaching experience and were doing their graduate studies at the time of the study. Paul, the only native speaker of English, was a novice to the local context (his first term teaching in Tunisia) but had considerable experience teaching English in an ESL context.

2.2. *Writing tasks*

Two argumentative topics similar to topics usually assigned in EFL writing exams at the participants' university were selected for the study. One topic asked students to compare and contrast the positions of being the oldest and youngest child in a family using their own experience (family topic). The second topic asked the students whether they agree or disagree that "technology has made the world a better place to live" (technology topic). Each student wrote an essay on the family topic and, three days later, wrote an essay on the technology topic under exam-like conditions.

2.3. *Rating scales*

For the purposes of this study, I identified four multiple-trait rating scales (Brown & Bailey, 1984; Hamp-Lyons & Henning, 1991; Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughhey, 1981; Weir, 1993) and four holistic scales (Educational Testing Service, 2000; Hughes, 1989; Song & Caruso, 1996; Tyndall & Kenyon, 1996) from the literature that I thought were appropriate for marking argumentative essays by EFL university students based on my own experience teaching EFL writing. I then asked seven EFL writing teachers with five to twenty years experience teaching and assessing EFL writing to (a) choose from each set two scales that they found most suitable for evaluating EFL university students' argumentative writing and (b) make suggestions to improve the scales. There was a general agreement on the *EFL Placement Test* seven-level scale developed by Tyndall and Kenyon (1996) as the best (by 5 teachers) or second best (by 2 teachers) holistic scale. The *Composition Grading Scale* (Brown & Bailey, 1984) was selected as the best (by 4 teachers) or second best (by 3 teachers) multiple-trait rating scale. This scale has five levels and five rating dimensions (content, organization, grammar, mechanics, and style). Based on the teachers' suggestions, I made few minor changes to the versions of the scales for this study.

2.4. *Procedures*

I randomly selected two sets of four essays (2 topics × 2 students) for the think-aloud protocol sessions (see below). The remaining 24 essays were rated silently. Each participating teacher

completed a background questionnaire concerning their language, education, and EFL teaching and rating experience. I then met each teacher individually to introduce him/her to and discuss the holistic scale and a Report Form attached to each essay. The Report Form asked the rater to assign a score to the essay, explain the score they assigned, and report any comments they might have concerning the rating scale. Except for discussing the content of the rating scale, no formal rater training was provided (cf. Hamp-Lyons & Henning, 1991). This was so for three reasons: (a) to simulate the rating procedures these teachers were used to, (b) to avoid affecting the way they approached the rating task, and (c) the difficulty of gathering the raters together for formal training. The teachers were encouraged, however, to report on any problems they might face while rating the essays. Each rater was then given a set of 24 essays (two topics × 12 students) in a unique random order and a Report Form attached to each essay. They were asked to rate the essays at home using the holistic scale. The four participants rated the essays over different periods of time.

Next, each teacher was trained on thinking aloud and then asked to rate a new set of four essays (2 topics × 2 students) while thinking aloud. For this task, the teachers were asked (a) to verbalize their thoughts and feelings about the essays, the topics, and the rating scale, and (b) to describe whatever they did (e.g., refer to scale or topics) while rating each essay as naturally as possible. The main goal of the think-aloud protocols was to collect real-time data on the participants' interpretations, uses of, and reactions to the rating scale, how they read and judged the essays, their decision-making behaviors, and the essay aspects they attended to (Cumming et al., 2002; Lumley, 2002; Milanovic et al., 1996; Wolfe, Kao, & Ranney, 1998). Immediately after the think-aloud session, each rater was interviewed about his/her use of and reaction to the rating scale and the problems s/he faced (cf. Brown & Bailey, 1984).

About three weeks later, I met each teacher to discuss the multiple-trait rating scale and a new Report Form. The teacher then rated the same set of 24 essays in a different random order of essays and topics using the multiple-trait rating scale and filled in a Report Form for each essay. Next, the teacher rated a second set of four essays while verbalizing his/her thoughts. Finally, the teacher was interviewed about his/her use of and reactions to the multiple-trait scale and his/her opinion concerning the two rating scales.

2.5. *Data analysis*

Generalizability (G-) theory was employed to examine the effect of the two rating scales on the dependability of essay scores. G-theory provides a theoretical framework and a set of procedures for estimating the relative effects of different factors (called facets) on test scores in performance assessment (Bachman, 1997). This process consists of two stages: a Generalizability study (G-study) and a Decision study (D-study). The goal of the G-study is to identify and quantify the sources of variance in test scores attributed to each facet (student, task, rater, etc.) in the testing setting. This information is then used to estimate G-coefficients and dependability (ϕ) indices and to examine the relative effects of varying the number of conditions in each facet (e.g., number of raters and/or tasks) on the dependability of scores in the D-study (Bachman, 1997; Brennan, 2001; Shavelson & Webb, 1991). G-coefficients are calculated with an error term that reflects how well the observed scores differentiate the test-takers on the ability being measured (i.e., norm-referenced [NR]). Dependability (ϕ) indices, on the other hand, are calculated with an absolute error term and indicate the degree of dependability that exists for an observed score as representing the individual test-taker's standing in relation to a well-defined criterion or domain of ability (i.e., criterion-referenced [CR]) (Lynch & McNamara, 1998, p. 167).

I used the computer program *GENOVA* (Crick & Brennan, 1984) to estimate (a) the relative contributions of students, topics, raters, and their interactions to variance in essay scores for each rating scale, and (b) the dependability of scores with various numbers of topics and raters for the holistic and multiple-trait scales. In this study, students were the object of measurement, while topic and rater were considered random facets in the sense that the four raters and two topics were considered interchangeable with any other set of four raters and two topics from the universe of admissible observations (Lynch & McNamara, 1998). The rating scales were treated as a fixed facet, as the scales are not generalized beyond the conditions sampled in the G-study. Following a suggestion by Shavelson and Webb (1991, p. 77) concerning fixed facets, a G-study was conducted on (a) each of the five conditions of the fixed facet (i.e., each multiple-trait scale), and (b) the average score across the five multiple-trait scales. Since all the students (S) wrote essays on both topics (T) and were scored by all the raters (R), the G-study design was S × T × R (i.e., completely crossed).[1] Because I examined scores assigned with reference to specific rating criteria, I report only dependability coefficients (i.e., criterion-referenced interpretations) below. It should be noted, however, that dependability (Φ) indices are often smaller than G-coefficients because while the former decrease as both the main effects of the facets (e.g., task, rater) and interactions between them and the object of measurement (i.e., examinee) increase, the latter decrease only if the facets interact with (i.e., affect) the object of measurement (Lynch & McNamara, 1998).

I transcribed and segmented the verbal protocols into separate decision-making statements using the following criteria from Cumming et al. (2002): pauses of five seconds or more, rater reading aloud a segment of the essay, and/or end or beginning of the assessment of a single essay. Although reading the rating scale descriptors could be used to indicate decision-statement boundaries, I felt that these units should be included in the analysis in order to find out how the raters used the rating scales. I used Cumming et al.'s (2002, p. 77) empirically based scheme of rater decision-making behavior to code the think-aloud protocols and data from the Report Forms. Cumming et al.'s model consists of 35 decision-making behaviors grouped under three foci and two strategies. The three foci are: self-monitoring (i.e., focus on one's own rating behavior), the essay's realization of ideational and rhetorical elements, and the essay's accuracy and fluency in the English language. In terms of strategies, the model includes interpretation strategies, i.e., reading strategies aimed at comprehending the essay, and judgment or evaluation strategies for formulating ratings or scores. According to this model, essay rating is an interactive process wherein the rater reads, judges, exercises diverse self-control strategies, and attends to numerous aspects of essays simultaneously.[2]

Each decision-making statement in the think-aloud protocols was assigned all relevant codes in the coding scheme. Data obtained through the Report Forms were also coded in terms of the essay aspects the raters mentioned, using categories under the language-judgment and rhetoric-and-ideas-judgment strategies in Cumming et al.'s (2002) scheme. I coded all the qualitative data in this study. To check the reliability of the coding, I discussed the coding scheme with another

[1] While the sample size is small, it seems sufficient for this statistical analysis. Swartz et al. (1999), for example, conducted similar analyses with a slightly smaller sample (20 students and 4 raters).

[2] I made two changes to the coding scheme for the purposes of this study. First, a new category, "read, interpret, refer, or comment on rating scale" was added under interpretation strategies and self-monitoring focus to account for the raters' uses of the rating scales. Second, to find out the source of the rating criteria and the scoring decisions the raters employed, the categories "Define or revise rating criteria" and "Articulate, justify or revise scoring decision" (strategies 8 and 11 in the Appendix) were further analyzed using Wolfe et al.'s (1998) distinction among self-generated, essay-comparison, and scale-based criteria.

researcher, who was doing an MA in TESL at the time of the study, who then independently coded random samples of 15% of the think-aloud protocols and the Report Forms. Percentage agreements achieved were 86% for the coding of the Report Form data and 75% for the think-aloud protocols. Because the raters did not provide the same amount of data in both think-aloud protocols and Report Forms, percentages were computed for each category and analyses were run on the percentages. Use of percentage counteracts the length of protocols as an influence and puts the comparisons (within and across raters) on an equivalent basis. I analyzed the interviews qualitatively to identify the raters' perceptions and the problems they faced in using the scales.

3. Findings

3.1. Rating scale effects on essay scores

Using the variance components from the G-study, I conducted a series of D-studies to investigate the relative effects of varying the numbers of topics and raters for the holistic and multiple-trait scoring methods. Table 2 reports the variance components for the D-study for the same sample sizes used to estimate the G-study variance components (i.e., four raters and two topics) to obtain the relative effects as well as the reliability and dependability estimates for the original data set. The results in Table 2 reflect the relative effects of raters and topics on the holistic and multiple-trait scores collected in this study. For the holistic scale, most of the variance (69%) is attributable to students, or universe score variance, a desirable result since the primary purpose of the assessment is to differentiate the writing abilities of the students. However, the student-by-topic-by-rater (STR) interaction also contributed substantially (20%) to the total variance on the holistic scale indicating that a large error variance is not explained by the design in this study. In addition, the relatively large student-by-topic (ST) interaction (7%) suggests that certain students performed better on one topic than the other. Finally, Table 2 shows that the rater facet accounted for 4% of the variance, meaning that the four raters applied slightly different standards when rating the essays holistically.

For the individual multiple-trait scales, the greatest percentage of variance is attributable to the rater facet, particularly for the style scale (74%), while for organization, rater contributed 37% of the variance. This indicates that certain raters were more lenient or severe than others across all students on these rating dimensions. Comparing the rater variance component to the variance component for students, there was a relatively small effect for student ability on the variability of the multiple-trait scores (between 6% and 45%). The STR interaction accounted for 15% (for mechanics) to 28% (for organization) of the total variance, indicating a large unexplained error variance. The small variance component of the facet topic (0% for style to 2% for grammar) indicates that this facet had a very small effect on the essay multiple-trait score variability. The student-by-topic (ST) interaction accounted for 0% (for style) to 10% (for organization). The high ST interaction for organization indicates that certain students performed better on one topic on this rating dimension. Finally, the RT interaction contributed no variance, except for the mechanics scale (1%). When scores are averaged across the five multiple-trait scales, the rater facet contributed most of the variance (62%), followed by the STR interaction (20%) and student (18%).

These results indicate different patterns of interactions among raters, students, and topics depending on the rating scale being used. For example, the student-by-topic interaction was high for the holistic and organization scores but low for the other multiple-trait scales. The high student-by-topic interaction for the holistic and organization scales indicates that the students

Table 2
Variance components for D-study S × R × T (4 raters and 2 topics)

Source	Student (S)			Topic (T)			Rater (R)			Student-by-topic (ST)			Student-by-rater (SR)			Topic-by-rater (TR)			Student-by-topic-by-rater (STR)			Total var.	Ep2 (NRT)	Φ (CRT)
Scale	VC	*SE*	%	VC	*SE*	%	VC	*SE*	%	VC	*SE*	%	VC	*SE*	%	VC	*SE*	%	VC	*SE*	%			
Holistic	.293	.156	68.9	.000	.006	0.0	.016	.016	3.8	.028	.046	6.6	.000	.023	0.0	.002	.005	0.5	.086	.019	20.2	.425	.72	.69
Content	.038	.041	20.8	.000	.002	0.0	.086	.057	47.0	.003	.017	1.6	.017	.015	1.6	.000	.001	0.0	.039	.009	21.3	.183	.39	.21
Organization	.049	.053	25.9	.000	.002	0.0	.069	.046	36.5	.018	.029	9.5	.000	.014	0.0	.000	.001	0.0	.053	.012	28.0	.189	.41	.26
Grammar	.015	.025	9.5	.003	.005	1.9	.103	.067	65.2	.003	.015	1.9	.001	.010	0.6	.000	.002	0.0	.033	.007	20.9	.158	.29	.10
Mechanics	.178	.037	45.2	.000	.006	0.0	.155	.103	39.3	.000	.022	0.0	.000	.016	0.0	.003	.005	0.8	.058	.013	14.7	.394	.23	.08
Style	.012	.023	6.5	.000	.001	0.0	.138	.089	74.2	.000	.013	0.0	.000	.011	0.0	.000	.001	0.0	.036	.008	19.4	.186	.25	.06
Average MT	.028	.025	18.1	.000	.001	0.0	.096	.062	61.9	.000	.012	0.0	.000	.009	0.0	.000	.001	0.0	.031	.007	20.0	.155	.47	.18

VC, variance components; SE, standard error; %, percentage of total variance; Φ, dependability coefficient (CRT); Ep2, reliability coefficient (NRT); MT, multiple-trait scale.

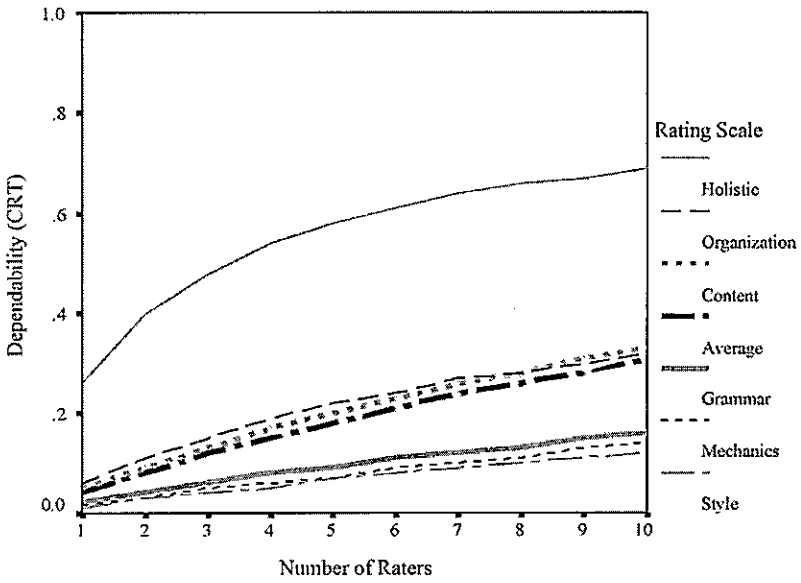

Fig. 1. Dependability for 1 topic.

were ranked differently on these two rating scales depending on the topic assigned. Although rater variability was high for all scales, the results indicate that it was higher for the multiple-trait scales – particularly for style (74%), grammar (65%), and content (47%) – than for the holistic scale (4%).

Table 2 reports the dependability indices (Φ) for each rating scale. Dependability indices indicate the degree of dependability that exists for an observed score as representing the individual test-taker's standing in relation to a well-defined criterion or domain of ability (i.e., criterion-referenced). Table 2 indicates that the holistic scale resulted in higher Φ (and G-) coefficients than the individual and averaged multiple-trait scores. The organization and content scales obtained higher dependability (and G-) coefficients than the three language scales (grammar, mechanics, and style), which resulted in very low coefficients. Averaging across the five multiple-trait scales resulted in a slightly higher G-coefficient (.47) but a lower dependability index (.18).

I used the G-study variance component to estimate dependability indices (CRT-Dependability) for one D-study design with one to ten raters and one to four topics. The D-study design discussed here is: all students take all topics, which are scored by all raters (i.e., completely crossed: $S \times T \times R$).[3] For all designs and combinations of raters and topics, the style, grammar and mechanics scales resulted in low dependability coefficients. Figs. 1–4 graph the dependability (Φ) indices obtained for the completely crossed design. They indicate that as the number of topics and, particularly, raters increases, score reliability increases. While holistic scoring seems to achieve acceptable Φ indices (i.e., above .80), albeit with large numbers of tasks and raters (e.g., 3 topics and 6 raters or 4 topics and 4 raters per student), the multiple-rating scales, particularly

[3] I examined other designs as well (e.g., rater nested within topic [$S \times (R:T)$] and student nested within rater, rater nested within topic [S:R:T]), but these designs resulted in lower coefficients. For a detailed technical explanation on how to conduct a D-study to examine the relative effects of varying the number of conditions in each facet (e.g., number of raters and/or tasks) on the dependability of scores, see Brennan (2001) and Shavelson and Webb (1991).

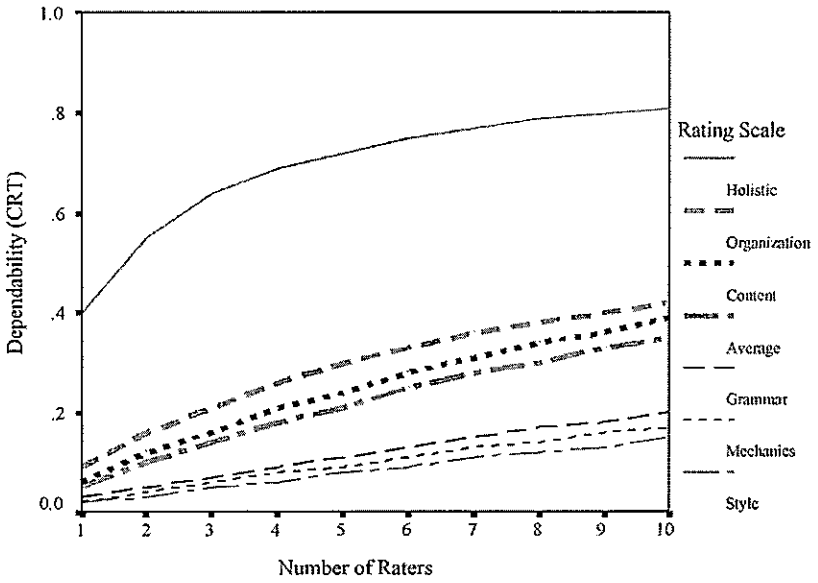

Fig. 2. Dependability for 2 topics.

style, grammar and mechanics, seem to require much more ratings to achieve acceptable score reliabilities (cf. Swartz et al., 1999).

ANOVA and post-hoc Tukey HSD tests detected significant differences between all raters on all the scales. Fadwa assigned significantly lower scores on all the scales, whereas Hatem assigned significantly lower holistic scores but significantly higher scores on the five multiple-trait scales.

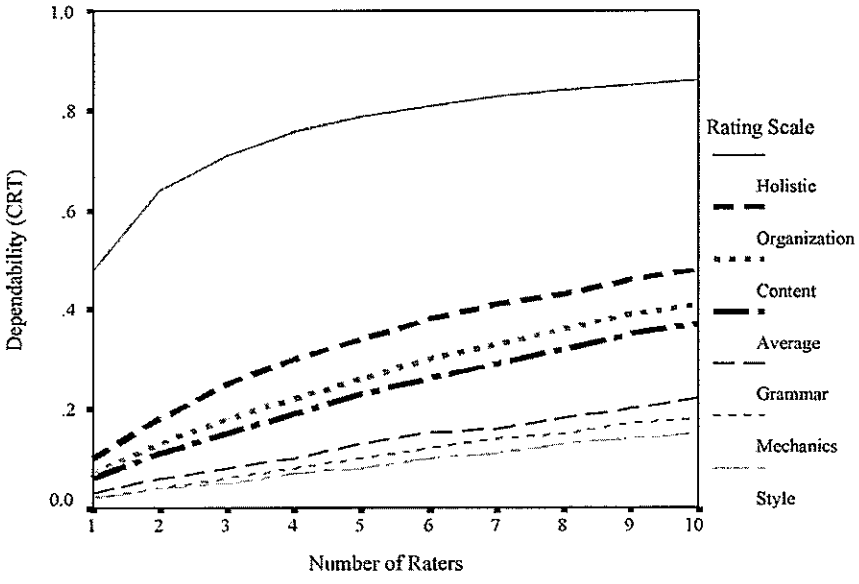

Fig. 3. Dependability for 3 topics.

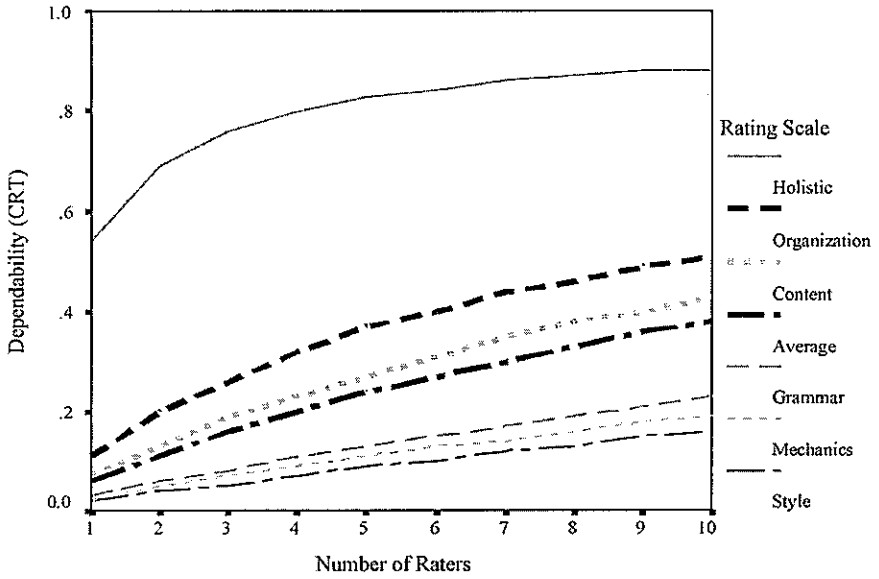

Fig. 4. Dependability for 4 topics.

Paul assigned higher scores on all the scales except for style, while Nader assigned scores in the middle on all the scales. These results are presented in Table 3. If we take these differences as a crude measure of rater leniency/severity, they suggest a significant rater-rating scale interaction effect.[4]

In sum, the G-theory and ANOVA analyses indicate that the four teachers assigned different scores and ordered the essays in different ways in terms of the five multiple-trait scales, particularly for the content and language scales (grammar, style, and mechanics). The holistic and, to a lesser extent, organization scores, in contrast, show a relatively higher level of score reliability. However, these two scales resulted in the highest student-by-topic interactions as well. The D-study results indicate that to increase the dependability indices, a large number of raters and topics need to be included, particularly for the multiple-trait scales. Furthermore, the ANOVA results suggest that there might be significant rater-by-rating scale interactions. These findings will be discussed further following discussion of the effects of rating scales on teachers' rating processes.

3.2. Rating scale effects on essay rating processes

Thirty-two think-aloud protocols were collected for this study (4 raters × 4 essays × 2 rating scales). However, because of technical problems, two of the holistic protocols were lost. The results reported below concern thirty think-aloud protocols. The average number of decision-making statements coded in the verbal protocols was 27. However, the protocols varied in length across raters and rating scales. For example, Fadwa produced 36 decision-making statements on average, while Hatem produced 21 statements. In addition, the raters produced more decision-making statements ($M = 36$) when rating the essays holistically than when using the multiple-trait scale ($M = 20$).

[4] Keep in mind, however, that these raters did not receive any training, standardizing, or feedback.

Table 3
Means and standard deviations of scores assigned by four raters

Scale	High mean scores ◄								► Low mean scores		
	Rater	*M*	*SD*	Rater	*M*	*SD*	Rater	*M*	*SD*	Rater	*M*	*SD*
Holistic*	Paul	4.34	.95	Nader	3.89	1.13	Hatem	3.59	.96	Fadwa	3.57	.75
Content	Paul	3.67	.62	Hatem	3.41	.59	Nader	2.71	.74	Fadwa	2.38	.69
Organization	Paul	3.52	.75	Hatem	3.44	.53	Nader	2.59	.72	Fadwa	2.48	.66
Grammar	Hatem	3.44	.43	Paul	2.76	.60	Nader	2.75	.48	Fadwa	1.83	.71
Mechanics	Hatem	3.60	.44	Paul	3.15	.69	Nader	2.68	.52	Fadwa	1.72	.93
Style	Hatem	3.12	.49	Nader	2.64	.45	Paul	2.42	.49	Fadwa	1.42	.69

* Holistic scale: 7 points; multiple-trait scales: 5 points.

Table 4
Percentage means and standard deviations of essay features in Report Forms (4 raters and 24 essays)

	Holistic		Multiple-trait		Total	
	M	*SD*	*M*	*SD*	*M*	*SD*
General comment	4.25	3.83	9.10	9.93	6.67	7.69
Reasoning, topic development	13.86	6.49	10.55	6.46	12.20	6.49
Task completion	12.13	9.16	7.96	10.26	10.05	9.64
Relevance	5.44	5.34	1.30	2.03	3.37	4.45
Coherence	6.74	6.15	3.00	4.36	4.87	5.50
Interest, originality or creativity	.24	.68	.00	.00	.12	.48
Redundancies	.37	1.04	.00	.00	.18	.74
Text organization	15.07	7.10	17.37	12.59	16.22	9.95
Style, register or genre	4.38	3.29	8.39	6.37	6.39	5.32
Ideas or rhetoric	5.32	5.38	4.71	4.08	5.01	4.62
Layout	5.98	6.31	1.86	2.36	3.92	5.07
Quantity of total written production	.48	1.36	1.66	2.62	1.07	2.10
Comprehensibility	.74	1.23	.57	1.07	.65	1.12
Gravity of errors	.50	1.41	.00	.00	.25	1.00
Error frequency	3.62	3.34	2.89	4.66	3.25	3.93
Fluency	5.19	3.75	1.69	2.04	3.44	3.43
Lexis	.00	.00	.00	.00	.00	.00
Syntax or morphology	3.77	2.61	1.25	2.32	2.51	2.72
Spelling or punctuation	7.81	4.23	7.26	7.77	7.53	6.05
Language overall	1.30	1.40	5.78	7.27	3.54	5.56
Total rhetoric and ideas	63.55	14.05	53.28	27.25	58.42	21.60
Total language	29.38	8.09	22.96	17.32	26.17	13.48

Table 4 compares the relative frequencies of the aspects of essays mentioned in the Report Form across the two rating scales. The holistic scale prompted more comments on reasoning, logic, or topic development, task completion, relevance, coherence, layout, fluency, and syntax or morphology, while the multiple-trait scale resulted in relatively more comments on style and register, text organization, and language overall. The raters made more comments on the essays overall when using the multiple-trait scale compared to the holistic scale. These were comments that could not be classified as either pertaining to the rhetorical or linguistic aspects of the essays. As a result, there were fewer comments on rhetoric and language for multiple-trait rating than for holistic rating. This was surprising because I had assumed that the multiple-trait scale would lead raters to make more specific comments. Note also that most categories showed large variability as indicated by the standard deviations in Table 4.

The frequencies of the 35 decision-making behaviors of the four raters across rating scales are reported in the Appendix. These results show that the rating scales did not markedly affect the rating processes of the raters nor the essay aspects they attended to while rating. Except for the category "Read or reread composition" which was more frequent in the holistic ratings, there were no significant differences in the frequencies of the 35 strategies across rating scales. The multiple-trait scale resulted in a slightly higher frequency of the strategies "read or interpret rating scale," "summarize, distinguish or tally judgments collectively," and "articulate, justify, or revise scoring decision." Note also that the holistic scale elicited slightly more "interpret ambiguous or unclear phrase" statements, while the multiple-trait scale led to slightly more "assess style, register or genre" and "consider gravity of error" statements. In terms of the source of the rating criteria and

Table 5
Mean percentages and standard deviations of sources of rating criteria

	Holistic (14 protocols)		Multiple-trait (16 protocols)		Total (30 protocols)	
	M	*SD*	*M*	*SD*	*M*	*SD*
Define or revise rating criteria*						
Scale-based	53.03	37.13	79.52	35.90	67.87	38.11
Self-generated	50.00	45.34	25.83	32.70	36.47	39.82
Essay-comparison	12.12	22.47	5.95	15.48	8.67	18.71
Articulate, justify, or revise scoring decision*						
Scale-based	65.82	36.12	91.62	21.73	79.58	31.63
Self-generated	34.52	41.07	23.03	19.83	28.81	36.71
Essay-comparison	12.76	29.14	3.13	12.50	7.62	22.03

* Percentage of each subcategory is based on the total of its respective category.

the scoring decision the raters employed, Table 5 shows that the raters used scale-based criteria more often, particularly with the multiple-trait scale, but relied relatively more frequently on essay-comparison when rating the essays holistically. Moreover, the participants used relatively more self-generated rating criteria when rating the essays holistically.

Table 6 reports the frequency of the overall categories of decision making the raters employed. The multiple-trait scale resulted in more judgment strategies, while the holistic scale prompted more interpretation strategies. This was an expected result since the raters had to make more than one score decision with the multiple-trait scale. When using this scale, the four raters produced more judgment decision statements (e.g., read scale, revise criteria, and articulate/revise scores). In terms of focus, the scales resulted in about an equal amount of attention to self-monitoring, rhetoric and ideas, and language. However, the multiple-trait scale prompted relatively more self-monitoring and language judgment strategies, while the holistic scale prompted more self-monitoring interpretation strategies. Both scales prompted an equal amount of attention to judgment strategies about rhetoric and ideas. These results suggest that the content and organization of the rating scales in this study did not much affect the essay aspects raters attended to, nor the frequency with which they considered these aspects.

Table 6
Distribution of two types of strategies and three types of focus across rating scales

	Holistic (14 protocols)		Multiple-trait (16 protocols)		Total (30 protocols)	
	M	*SD*	*M*	*SD*	*M*	*SD*
Interpretation strategies	51.74	8.38	40.98	8.66	46.00	10.01
Judgment strategies	49.04	9.48	60.33	9.97	55.06	11.16
Self-monitoring focus	37.46	3.64	37.43	6.64	37.45	5.36
Rhetoric and ideas focus	34.57	9.05	33.17	7.64	33.83	8.21
Language focus	30.22	11.60	32.75	12.96	31.57	12.20
Self-monitoring/interpretation	30.24	4.20	23.45	3.01	26.62	4.94
Self-monitoring/judgment	7.23	3.57	13.98	7.64	10.83	6.90
Rhetoric and ideas/interpretation	10.29	4.40	8.03	3.12	9.08	3.88
Rhetoric and ideas/judgment	24.28	9.85	25.15	8.14	24.74	8.83
Language/interpretation	11.39	4.60	9.65	5.67	10.46	5.19
Language/judgment	18.83	7.91	23.09	8.34	21.11	8.29

Three factors, topics, raters, and essays, seem to have contributed to the variability in the mean percentages observed in Tables 4–6. Concerning topic effects on the rating processes of the four teachers, the two topics prompted the same amount of comments on rhetoric (58%) and language (26%), but the family topic resulted in more comments on task completion (12%) and fewer on text organization (13%) than did the technology topic (8% and 19%, respectively) in the Report Forms. The two topics did not affect the frequency of the types of strategies and focuses in the think-aloud protocols. There was much variability, however, across raters in terms of decision-making behaviors and essay aspects attended to as reported in the think-aloud protocols and the Report Forms. Given the very small sample size of teachers and essays in this study, it is difficult to interpret these findings.

3.3. Individual rating styles

An impressionistic analysis of the think-aloud protocols indicated that the rating processes the four teachers employed differed across raters but not across rating scales. While rating an essay, Hatem would compare it to what he expected and then articulate his preferences. He made relatively many comments on the essay writers in terms of their linguistic competence and attendance of writing classes as well as on essay style, particularly in terms of the distinction between written and spoken modes. When rating the essays holistically, Paul assigned a provisional score to the essay after reading the first paragraph and then modified the score in light of what followed. This provisional score concerned the essay's rhetorical aspects (first two points in the holistic scale), while the revision was based on language factors. With the multiple-trait rating scale, Paul read the essay and then assigned a score to each aspect while commenting on the essay in terms of the specific category under focus. Paul seems to have found it easier to separate the rating criteria into two groups (rhetoric and language) when rating the essays holistically. Nader compared essays to each other, an approach he admitted using when rating exam essays. Fadwa made relatively many predictions about the organization and content of the following paragraphs and suggestions for alternative ways to develop the essay while reading an essay. Most notably, Fadwa explained most of the scores she assigned with reference to her own reaction to the essay, rather than with reference to the rating scales (e.g., "I liked..."; "What makes me feel uncomfortable is..."). The use of these 'internal criteria' might explain the discrepancies between the scores the participants assigned. A common behavior among all the raters was their expression of their preferences for how the essays should have been written and referring to what they had taught while reading and evaluating the essays. Although they were explicitly instructed to evaluate the essays with reference to the rating scales provided, the four teachers in this study could not avoid referring to what they teach and expect when rating the essays for this study.

3.4. Raters' perceptions of the rating scales

The four teachers found the rating criteria in both scales congruent with the objectives of the EFL writing course they were teaching at the time of data collection. They further believed that the criteria were appropriate for evaluating essays on argumentative topics like the ones included in this study. They had an overall positive attitude towards both rating scales and found the level divisions (and rating dimensions for the multiple-trait scale) discrete and the descriptors clear and appropriate. For the holistic scale the four participants reported that they found it sometimes difficult to distinguish between levels (e.g., 5 vs. 6), while Paul commented

that holistic rating does not provide clear guidance on how to determine a final score when an essay displays different levels of proficiency at the linguistic and rhetorical levels. As a result, this scale often "leads to conflicting scores" in terms of these two dimensions and, by urging the rater to make a decision on a final score, increases the cognitive load on the rater.

Paul reported that he did not face this problem with the multiple-trait scale since he did not have to synthesize or decide on a final score for each essay. Nader also commented that because the multiple-trait scale "considers each aspect on its own" it is "fairer to the student and effective for the teacher" while Fadwa described this scale as being more "scientific," "fair," and "honest." All raters agreed, however, that multiple-trait scoring took longer and made the rating process slower compared to holistic rating since the rater has to consider five rating dimensions simultaneously. But the participants commented that with more practice, it would take them less time to rate the essays using either rating scale. Finally, Hatem pointed out that achieving a high score on content seemed very easy compared to other dimensions, particularly style. Unlike the other three participants, Hatem preferred the holistic scale, which he found "better," "more practical," and "familiar" (compared to impressionistic rating).

4. Summary and discussion

Contrary to the hypothesis that multiple-trait scoring will result in higher score reliability (Goulden, 1992), a higher level of score reliability was achieved when the essays were marked holistically. The multiple-trait scales resulted in high rater variability and required more ratings to achieve acceptable dependability indices. It is worth noting also that the rhetorical scales (content and organization) led to relatively higher score reliability than did the three linguistic scales (grammar, mechanics, and style).

One possible explanation for the relatively higher score reliability for holistic rating is that the participating raters relied on the impressionistic criteria they are used to when rating the essays holistically. This high score reliability, however, might be hiding discrepancies between the teachers' evaluation criteria. The multiple-trait scale seems to expose these differences. Nevertheless, while the raters might have employed other criteria than those I provided when rating the essays holistically, the supplemental aspects that they employed led to a relatively higher level of agreement, contrary to the hypothesis that supplementary aspects may lead to discrepancies across raters (Goulden, 1992, 1994). These "supplementary criteria" are reflected in the higher frequency of self-generated and essay-comparison criteria in the holistic think-aloud protocols. This might also be true for the relatively higher score reliability achieved when assessing essay organization, since text organization constituted the main focus of the writing course the four participants were teaching at the time of data collection. This interpretation finds support in Erdosy's (2004) finding that teachers with similar teaching experiences tend to have similar views of the nature of second language proficiency and, as a result, "are likely to base their judgments on a shared construct of writing proficiency" (p. 57).

It is worth noting here that the dependability indices obtained in the current study are somewhat lower than those obtained in previous studies that used G-theory to evaluate rater and task effects on essay scores (e.g., Lee & Kantor, 2005; Schaeffer, Briel, & Fowles, 2001). For example, while Fig. 2 reports a reliability of .55 for two topics and two raters (holistic rating), Lee and Kantor obtained a figure of .70. Moreover, Lee and Kantor found that task-related variability was larger than that due to rater-related effects. As a result, a larger number of tasks, but fewer raters, were

needed to achieve acceptable values of score reliability in their study (cf. Schaeffer et al., 2001). Lee and Kantor employed certified raters and used careful reader training, however, whereas the current study did not. That would appear to be the main reason for the lower score reliabilities obtained in the present study.[5]

There are two other observations worth noting concerning rating scale effects on essay rating. First, there were significant student-by-topic interactions for the holistic and organization scales, but not for the other scales. Second, there were significant rater-by-rating scale interaction effects for some raters whose leniency/severity, compared to other raters, seemed to depend on the rating scale and dimension being assessed. Future studies need to consider these interaction effects.

The think-aloud data suggest that the content and organization of the rating scales did not much affect the essay aspects the raters in this study attended to, nor the frequency with which they considered these aspects. The rating scales seem to have four slight effects on the decision-making behaviors of the raters in this study, however. First, the holistic scale prompted more decision-making statements than did the multiple-trait scale for all the raters, which might suggest that the raters had to verbalize, or explain, more often when rating the essays holistically. Second, the multiple-trait scale resulted in more judgment strategies while the holistic scale prompted more interpretation strategies. While this was expected since the raters had to make more than one score decision with the multiple-trait scale, it indicates that the way the rating criteria are organized in an evaluation scheme affects the relative frequency of the strategies raters use. Third, while the raters in this study used scale-based criteria relatively more frequently than essay-comparison and self-generated rating criteria, they did so more often with the multiple-trait scale. The holistic scale resulted in relatively higher frequency of essay-comparison and self-generated criteria. As pointed out above, these "supplementary criteria," however, seem to have led to a higher inter-rater agreement. Finally, unlike the other three participants, Paul, the least experienced rater, felt a need to separate the holistic scale into two dimensions, rhetoric and linguistic, to make the rating task more manageable when rating the essays holistically. These findings suggest that there might be significant interaction effects between scales and raters on essay rating processes.

Raters were the major source of variability in terms of the scores assigned and the frequencies of the decision-making strategies used, particularly for multiple-trait rating. This variability is surprising, particularly for Hatem, Nader, and Fadwa who had been teaching the same writing course and rating essays together for more than six semesters, and, thus, were expected to have developed a relatively common set of criteria that they would have brought to the rating task when using the new scales. While this was true, to a certain extent, for the holistic and organization scores these raters assigned, the low score reliability coefficients reported above indicate that these raters brought conflicting criteria to the evaluation task. Given that the raters differ in many respects (e.g., teaching and rating experience, academic background), it is difficult to determine which rater variables affected the scores they assigned. Rater-rating scale interaction is another area for further research.

5. Conclusion

Do the findings of this study show any basis for choosing one rating scale over the other? While the findings of this study favor holistic scoring, the answer to this question depends on the purpose

[5] I would like to thank one of the anonymous reviewers for drawing my attention to these two studies and how their findings compare to the findings of the current study.

of assessment and its context. As Hamp-Lyons (1991) urged, designing and implementing a writing assessment is an iterative process that should include considerations about scoring procedures from the very beginning (p. 263). Rating scales will interact with other components of the assessment (e.g., tasks, raters) to strengthen or undermine the whole assessment system (Hamp-Lyons, 1990; Weigle, 2002). As such, it is crucial to involve raters in the process of developing or adapting rating scales for a particular assessment context in order to ensure that they appropriate the rating scale and use it appropriately and consistently (Davidson, 1991; Hamp-Lyons, 1991; Weigle, 2002). The use of multiple research methods can enhance this process and provide test developers and users with valuable information to improve the validity of the interpretations and fairness of decisions based on essay scores (Lynch, 1996).

Contrary to what was expected, the findings suggest that the holistic scale resulted in higher score reliability. The multiple-trait scale, on the other hand, resulted in lower score reliability and, as a result, seems to have little practical value in this context. However, this low score reliability might be due to the lack of rater training, which is a major limitation of the current study and gives an indication of what happens when rater training is not used. Rating scales are not the sole determinants of essay scores (Erdosy, 2004; Lumley, 2005). Even with specific, detailed rating scales, inter-rater reliability cannot be guaranteed, unless rating criteria are clarified and discussed extensively, anchor papers are established, and careful rater training is provided (Davidson, 1991; Erdosy, 2004).

Acknowledgement

I would like to thank the participating students and teachers as well as Alister Cumming, Fred Davidson, and Mohamed Daoud for comments and feedback on the design of this study and earlier drafts of this article.

Appendix A. Mean percentages and standard deviations for 35 decision making behaviors across rating scales

	Rating scale					
	Holistic (14 protocols)		Multiple-trait (16 protocols)		Total (30 protocols)	
	M	*SD*	*M*	*SD*	*M*	*SD*
Self-monitoring focus						
Interpretation strategies						
1. Read or interpret essay prompt	.88	.85	1.13	.84	1.01	.84
2. Read or reread composition	25.63	4.73	15.90	4.87	20.44	6.83
3. Envision personal situation of the writer	1.99	1.80	1.67	1.96	1.82	1.86
4. Scan whole composition	.27	.49	.65	.90	.47	.75
5. Read or interpret rating scale	1.47	.95	4.10	2.94	2.88	2.58
Judgment strategies						
6. Decide on macro-strategy for reading and rating	.46	.80	.16	.66	.30	.73
7. Consider own personal response or biases	1.30	1.30	1.89	1.76	1.62	1.57
8. Define or revise rating criteria	1.60	1.26	2.75	2.07	2.21	1.81

Appendix A (*Continued*)

	Rating scale					
	Holistic (14 protocols)		Multiple-trait (16 protocols)		Total (30 protocols)	
	M	*SD*	*M*	*SD*	*M*	*SD*
9. Summarize, distinguish, or tally judgments collectively	.88	.87	3.43	2.54	2.24	2.31
10. Articulate general impression	.54	.86	1.11	1.20	.85	1.08
11. Articulate, justify, or revise scoring decision	2.44	1.37	4.63	3.65	3.61	3.00
Rhetorical and ideational focus						
Interpretation strategies						
12. Interpret ambiguous or unclear phrase	4.80	3.52	1.93	1.83	3.27	3.06
13. Discern rhetorical structure	3.82	3.19	3.36	1.97	3.57	2.57
14. Summarize ideas or propositions	1.67	1.13	2.74	1.68	2.24	1.52
Judgment strategies						
15. Assess reasoning, logic or topic development	4.95	1.97	5.36	2.20	5.17	2.07
16. Assess task completion	2.54	3.49	2.05	1.32	2.28	2.53
17. Assess relevance	1.59	1.57	.61	.86	1.07	1.32
18. Assess coherence	2.24	2.62	1.21	1.03	1.69	1.97
19. Assess interest, originality or creativity	.77	1.05	.72	.96	.74	.99
20. Identify redundancies	.39	.62	.20	.46	.29	.54
21. Assess text organization	3.61	2.10	4.30	2.51	3.98	2.31
22. Assess style, register or genre	2.11	2.63	4.37	2.92	3.31	2.97
23. Rate ideas or rhetoric	6.08	3.85	6.33	2.96	6.21	3.35
Language focus						
Interpretation strategies						
24. Observe layout	2.32	1.77	1.77	1.78	2.03	1.76
25. Classify errors into types	6.58	3.72	5.96	4.44	6.25	4.06
26. Edit phrase for interpretation	2.49	2.15	1.92	1.97	2.19	2.04
Judgment strategies						
27. Assess quantity of total written production	1.05	1.22	1.41	1.45	1.24	1.34
28. Assess comprehensibility	2.26	2.77	3.35	2.82	2.84	2.80
29. Consider gravity of error	.74	1.04	2.85	1.93	1.86	1.88
30. Consider error frequency	.42	.79	1.37	1.74	.92	1.44
31. Assess fluency	.05	.17	.09	.37	.07	.29
32. Consider lexis	3.63	3.71	3.89	2.34	3.77	3.00
33. Consider syntax or morphology	4.69	2.90	4.40	2.59	4.53	2.70
34. Consider spelling or punctuation	4.96	2.77	3.63	2.15	4.25	2.50
35. Rate language overall	1.05	.72	2.11	1.39	1.61	1.23

References

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System, 29* (4), 371–383.

Bachman, L. F. (1997). Generalizability theory. In: D. Corson (Series Ed.) & C. Clapham (Vol. Ed.), *Encyclopaedia of language and education: Vol. 7. Language assessment* (pp. 255–262). Dordrecht, Netherlands: Kluwer.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning, 34* (4), 21–42.

Carr, N. (2000). A comparison of the effects of analytic and holistic composition in the context of composition tests. *Issues in Applied Linguistics, 11* (2), 207–241.

Cooper, C. R. (1977). Holistic evaluation of writing. In: C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 3–31). Urbana, IL: NCTE.

Crick, J. E., & Brennan, R. L. (1984). *GENOVA: A general-purpose analysis of variance system. Version 2.2*. Iowa City, IA: American College Testing Program.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7* (1), 31–51.

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

Davidson, F. (1991). Statistical support for training in ESL composition rating. In: L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 155–164). Norwood, NJ: Ablex.

Educational Testing Service. (2000). *TOEFL: Information bulletin for computer-based testing*. Princeton, NJ: Educational Testing Service.

Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions* (TOEFL Research Report RR-03-17). Princeton, NJ: Educational Testing Service.

Freedman, S. W. (1979). How characteristics of essay influence teachers' evaluation. *Journal of Educational Psychology, 71* (3), 328–338.

Goulden, N. R. (1992). Theory and vocabulary for communication assessments. *Communication Education, 41* (3), 258–269.

Goulden, N. R. (1994). Relationship of analytic and holistic methods to rater's scores for speeches. *The Journal of Research and Development in Education, 27*, 73–82.

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

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

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

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning, 41* (3), 337–373.

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000-writing: Composition, community and assessment* (TOEFL Monograph Series No 5). Princeton, NJ: Educational Testing Service.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughhey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MS: Newbury House.

Lee, Y., & Kantor, R. (2005). *Dependability of new ESL writing scores: Evaluating prototype tasks and alternative rating schemes* (TOEFL Research Report RR-05-14). Princeton, NJ: Educational Testing Service.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing, 19* (3), 246–276.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt: Peter Lang.

Lynch, B. K. (1996). *Language program evaluation: Theory and practice*. Cambridge: Cambridge University Press.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and Many-Facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15* (2), 158–180.

Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. In: M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Colloquium, Cambridge and Arnhem* (pp. 92–114). Cambridge: Cambridge University Press.

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

Schaeffer, G., Briel, J. B., & Fowles, M. E. (2001). *Psychometric evaluation of the new GRE writing assessment* (GRE Research Report RR-01-08). Princeton, NJ: Educational Testing Service.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22* (1), 1–30.

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing, 14* (2), 157–184.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

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

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., De Kruif, R. E. L., Reed, M., et al. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement, 59*, 492–506.

Sweedler-Brown, C. O. (1985). The influence of training and experience on holistic essay evaluation. *English Journal, 74* (5), 49–55.

Tyndall, B., & Kenyon, D. M. (1996). Validation of a new holistic rating scale using Rasch multifaceted analysis. In: A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 39–57). Clevedon: Multilingual Matters.

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

Weir, C. J. (1993). *Understanding and developing language tests*. London: Prentice Hall.

Wolfe, E. W., Kao, C., & Ranney, M. (1998). Cognitive differences in proficient and non-proficient essay scorers. *Written Communication, 15*, 465–492.