J. King Saud Univ., Vol. 12, *Lang. & Transl.*, pp. 1 - 22 (A. H. 1420 / 2000).

Examining the Analytic Marking Method: Developing and Using an Analytic Scoring Schema

Ibrahim Al-Fallay

Associate Professor, Department of English,
College of Arts, King Saud University,
Riyadh, Saudi Arabia

(Received A. H. 8/6/1419; accepted A. H. 10/2/1420)

Abstract. The purpose of this study is to investigate closely the analytic approach to scoring composition, and to examine its usefulness as a pedagogical tool. The data of this study came from three different sources. First, the scores of 55 Saudi Arabian students enrolled at the Intensive English Program of the English Department in the College of Arts of King Saud University, Riyadh, Saudi Arabia, on the writing, grammar, and vocabulary sections of the Intensive English Program Proficiency/Achievement Test (IEPPAT) were obtained. The second source of data was the scores of the same subjects on writing, reading, and grammar tests constructed by instructors of the Intensive English Program. and used as midterms and finals. Finally, the scores of the subjects of this study on the Department's courses taught in the semester following the Intensive English Program were also obtained. The findings of this study indicate that the analytic marking method has high intra- and inter-raters reliability indices. It also indicates that a scoring grid composed of the following six features or components: content, organization, vocabulary, grammar, punctuation, and spelling is appropriate and the number of points allocated to each feature is suitable. This study found that the features of vocabulary, spelling, and organization are easier for students to deal with than the features of punctuation, content, and grammar. It was also found that there were strong relationships among the various features. The features loaded on two different factors; one is more related to competency in the foreign language and the other is more associated with general knowledge of the world. It was also found that it is possible to predict current and future performance of students on tests measuring these features. Among the recommendations of this study is that using the analytic marking schema is invaluable since the scores obtained from this approach to composition scoring could be further utilized. It is also recommended that writing instructors help their students in developing compositions with effective content. Finally, it is suggested that writing instructors examine the scores which their students obtain in these features in order to discover their students' points of weakness and strength. Remedial exercises and practices should be developed to treat students' weakness and material developers should take these points into consideration when constructing writing materials, especially for intensive programs of English as a foreign language.

I. Introduction

It is unusual to find a foreign language program not devoting at least one course to the aim of sharpening the skill of writing, teaching students its strategies, and/or introducing them to various writing styles. It is also rare to find a test claiming to

assess students' proficiency in a foreign language without having a section to measure students' writing ability. The Educational Testing Service (ETS), developers of the Test of English as a Foreign Language (TOEFL) which is the most widely used proficiency test in the world, realized the importance of including a section to assess the skill of writing. Since 1986, the ETS had been including a section in the TOEFL, called the Test of Written English (TWE), with the aim of measuring the writing skill of students of English as a Foreign Language. The TWE test is given with the TOEFL as a required section at five administrations of the TOEFL twelve annual administrations (TOEFL Bulletin of Information [1].

Writing is considered the most difficult skill to master since it is affected by linguistic, psychological, and cognitive variables. In addition, the selection of many factors, such as the prompt to which students have to respond, plays a crucial role in reflecting their degree of the skill's mastery. But these are not the only factors that affect the accurate assessment of the proficiency in writing. The approach or method adopted for scoring composition adds also to the complexity of writing assessment. This study attempts to investigate closely one method of marking composition, namely the analytic method. The usefulness of the composition scores is usually limited to informing students about their scores on the writing test or to helping instructors to decide who is ready to be moved to the next level in the educational program. However, very invaluable information is missed. The aim of this study is to find out if it is possible to develop a scoring grid along the lines of previous research and to statistically investigate the appropriateness of selected features or components of that grid. It also aims at finding if there are strong relationships among these features and whether one single factor underlies these features. Finally, the possibility of using the scores of the analytic scoring method for pedagogical purposes is also examined. One of these purposes is to discover students' points of weakness and strength and to remedy any weakness students may have and to enforce their points of strength.

II. Review of Selected Literature

Generally, there are four approaches to scoring compositions: The holistic or impressionistic method, the error-count method, the primary trait scoring method, and the analytic method. Hadley describes the holistic or impressionist approach as one in which "one or more readers assigns a single grade (or rating) to a text based on an overall impression"[2, p. 343]. Hence, the criteria used in scoring composition may vary from one rater to another; and a composition's feature that might be of utmost importance to one rater may not enjoy the same level of significance given by another rater. In primary trait scoring, a single feature such as spelling, grammar, or content is scored holistically. This approach seems appropriate if the instructor is interested in investigating a certain feature. Finally and as the name implies in the error-count method, a point or more is deducted for every mistake a students makes. Sometimes, mistakes of certain type are given more weight than others. For example, a mistake in grammar may cause the deduction of two points whereas a mistake in spelling may lead to only one

point deduction. It is not obvious according to this method how the content of the composition may be scored. In the analytic method, the focus of this study, the composition is analyzed for several features, each is given a certain score, and the total score given to the composition is the sum of the scores of the various features. In this approach, there is an agreement among most raters on what should be counted as a feature. Among the most agreed upon features are grammar, mechanics, content, organization, fluency, and relevance; and they compose what is known as an analytic scheme. Zughoul and Kambal believed that the weight given to each feature should vary depending on the students' level of proficiency in the foreign language [3].

The analytic method has some drawbacks. According to Cooper the construction of a suitable analytic scheme is time-consuming [4]. To built a suitable scheme, six steps have to be followed. First, the features included in the scheme should be obtained from professional research and authentic samples of students' writing. Then, the scheme is tested in a pilot study. The quality of effective and ineffective features should be taken into consideration. Next, points should be assigned to each feature. Then, graders should use the new scheme with new samples of writing; and finally, the reliability estimates of the scheme should be computed. Perkins (1983) added another four disadvantages of the analytic scoring method. He outlined those as

> First, there is the problem of an immoderate standard … some graders may try to use an absolute standard of quality such as published professional writing. Second, the features to be analyzed are isolated from context and are scored separately … Third, the choice of categories can be vague and certainly arbitrary because the categories themselves are determined by the graders, who base their choices on a corpus of professional and student writing … The most serious drawback to analytic scales is that the scoring weight of a particular category must be adjusted for different kinds of discourse[5, p. 657].

The analytic scoring method has also some advantages. According to Hadley, the washback effect of this method is superior since it pinpoints to the instructor, the features that the students have difficulties in mastering, and hence suggests appropriate remedial exercises. Since it is hard for the classroom instructor to ask his colleagues to re-score his students' composition to ensure reliability, Heaton believed that using the analytic scoring method will help in overcoming this shortcoming [6]. The instructor had at his disposal a scoring scheme that informs him of the relevant features and the scores allotted to each feature. This will lead to high agreement among graders and eventually to high reliability of the method. Furthermore, Cooper deemed that the analytic scoring method is beneficial in program evaluation and for research purposes[4, p. 17]. For example, if the students obtain low scores on the parts of spelling or punctuation, this may indicate that the program of instruction lacks topics that teach or enforce these topics.

The consistency among raters analytically and holistically scoring the same composition was also investigated. Zughoul and Kambal for example, found a reliability index of .829 when the compositions of advanced students were analytically graded and .817 when they were holistically marked [3]. However, the reliability index of the holistic method was .813, which was higher than the reliability index of analytic marking of the composition of intermediate students. The same pattern was found also when the writing of beginners was marked by using both methods. The reliability of the holistic method was .851 whereas it was .778 for the analytic method. Similarly, Al-Fallay compared the two scoring methods [7]. Four EFL teachers were asked to mark the compositions of 52 intermediate students in an EFL program. The raters first graded the compositions holistically. A month later the same compositions were analytically graded. He reported a reliability index of .87 for the holistic methods and .83 for the analytic method. Furthermore, Cleary compared the reliability of the analytic and error-count methods and concluded that the reliability of the error-count method was higher that the reliability of the analytic method [8]. He took this as an indication of the superiority of the former method over the latter.

As part of the task of validating a multiple-choice cloze test (MCCT) to assess the writing proficiency of EFL learners, Al-Fallay calculated the Pearson product-moment correlation coefficient among the totals of his MCCT components and four features of an analytic scheme [7]. The components of the MCCT and the features of the analytic scheme had the same purpose: to assess students mastery of English grammar, vocabulary, mechanics, and unity and organization. The correlation ranged from .73, the correlation between the mechanics component of the MCCT and the mechanics feature of the analytic scheme, to .21, the correlation between the vocabulary component of the MCCT and the unity and organization feature of the analytic scheme. He stated that "the correlations of the subjects' scores on the M-C test's [MCCT] components with their scores on similar components [features] of the composition marked analytically are higher than the correlations among the dissimilar components of the M-C Test and the composition"[7, p. 14]. Although he took the aforementioned observation as an indication of the construct validity of the MCCT in assessing the writing proficiency of EFL learners, it is in fact an implicit support to the validity of the analytic method in composition scoring.

Al-Gusair sought to investigate the actual practices of teaching and evaluating composition at female public secondary schools in Saudi Arabia [9]. She surveyed 33 female high school teachers of English and 23 supervisors, in addition to attending some composition classes and analyzing samples of students' evaluated composition assignments and exams. She found that teachers and supervisors favored the use of the analytic method. However, in their actual practices they were using the holistic method. Their focus was more on the correct form than the flow of ideas. She also reported that "teachers never focused on the mechanics (indentation, punctuation, and capitalization)"[9, p. 152]. Among her observations were that the analytic method is superior to the holistic method and

that the analytic method should be adopted. Moreover, she suggested that the points devoted to each feature of the analytic method should be predetermined. To make students aware of their points of strength and weaknesses, she suggested that teachers should report the individual score given to each feature.

III. The Present Study

A. The Research Problem

Discussing the importance of teaching and shaping the skill of writing and the evaluation of composition any further is redundant. However, many studies in the literature have investigated the external consistency of the holistic method more than their investigation of the analytic method [5; 10; 11]. In addition, the analytic method has been accused of being a mere test of mechanics [12; 13]. Hence, this study aims at a comprehensive investigation of the analytic method of composition scoring. The consistency of the inter- and intra-raters using this method was investigated; in addition to the construct validity of the five features used in scoring composition: content, organization, grammar, spelling, and punctuation. Furthermore, the reliability of this method in predicting the future performance of students in tasks similar to these five features was also investigated. Finally, the reliability of the prediction of this method in future writing tasks was also investigated.

This study aims to answer the following questions:

1- Is it possible to develop and utilize an analytic marking schema that has both high intra- and inter-raters reliability indices?
2- What are the features or components of this marking schema?
3- What weight should be allocated to each feature? What are the justifications for such allocations of points among these features?
4- Will students' performance in these features be the same? Or will they find some features easier to deal with than other features? What are the features with which students do face difficulties?
5- Are these features related? If the answer is yes, what is the degree of their relatedness?
6- Is there a single factor underlying these features? Or are there more than one factor underlying them? What is/are the nature of this/these features?
7- Are there strong relationships between these features and similar tests that reflect students' current and future proficiency in the foreign language?
8- Is it possible to use the analytic marking scores for pedagogical purposes? How?

B. Hypotheses

The hypotheses of this study were nondirectional and the significance level was set at $p < .05$. The hypotheses are as follows:

H1: There are no significant differences between the means of subjects in the various features of the analytic marking schema.
H2: There are no statistically significant correlations among the subjects' means on the six features of the analytic marking schema.

H3: There are no statistically significant correlations among the subjects' means on the six features of the analytic marking schema and their means on corresponding teacher-made midterm and final examinations.

H4: The various features of the analytic marking schema will significantly load on different factors.

H5: There are no statistically significant relationships as measured by multiple and linear regression analysis between the six features of the analytic marking schema and the current and future performance of subjects in similar tasks.

C. Method

1. Subjects

The subjects of this study were 55 Saudi Arabian high school graduates enrolled at the English Department in the college of Arts of King Saud University, Riyadh, Saudi Arabia. Upon their acceptance in the Department, students usually have to enroll in an intensive English program where reading, writing, and grammar are taught unless they score 60 points out of 100 on the Department's entrance examination. The intensive English program lasts for one semester. In the eighth week, three teacher-made midterm examinations are administered to students of the intensive program. These examinations are geared at assessing students' command of reading, writing, and grammar. By the sixteenth week (the week of final examinations) similar tests are administered. Tests of both occasions are worth 40 points of the 100 point total. A week before the final examinations, the students' command of English language skills and grammar is assessed through a standardized test called the Intensive English Program Achievement/Proficiency Test (IEPPAT). This standardized test is worth 60 points. Students' scores on this test are added to their scores on the forty-point midterm and final examinations; and those who gain 60 points or more are allowed to register in the next semester, for the following seven courses: Eng. 111 (Writing Skills), Eng. 112 (Spoken English), Eng. 113 (Reading Comprehension), Eng. 116 (Remedial Grammar), Eng. 117 (Basic Skills), Eng. 119 (Listening), and Eng. 120 (Vocabulary Building), in addition to one course in Islamic culture. English skills are further shaped and refined through additional courses taught in following semesters. These courses focus on the four skills of English in addition to grammar. In their fifth semester, students have to choose between two paths: English literature or applied linguistics; and upon the successful completion of the courses of the eighth semester, students are granted a B.A. in the specialty they have selected.

2. Materials

In the first council meeting of the Department of English for the academic year 1997/1998, the council recommended that a standardized test be developed and administered to students in the intensive English program. The first use of the test was at the end of the first semester of the same academic year. The test consists of four parts: twenty multiple-choice items assessing listening comprehension, forty

multiple-choice items assessing the students' mastery of English grammar, twenty multiple-choice items assessing their knowledge of vocabulary, twenty multiple-choice items assessing reading comprehension, and two prompts to one of which students have to write a composition of at least twelve lines. Raters were instructed to mark the composition analytically. Scores were distributed among the six features as follows: twenty-five points to content, fifteen points to organization, ten points to vocabulary, twenty points to grammar, fifteen points to punctuation, and fifteen points to spelling.

The data of this study came from three different sources. First, the scores of the fifty-five students on part four of the IEPPAT administered in the first semester of the academic year 1997/1998 were tabulated with each feature occupying a separate column. Second, the scores of the subjects on the teacher-made midterm and final examinations were also considered. Finally, the scores of the same students on the midterm and final examinations of the seven courses taught in the second semester of the academic year 1997/1998 were also tabulated. It is noteworthy to mention that the number of students who took the IEPPAT in the first semester of 1997/1998 was 126 students. However, only the scores of students who passed the IEPPAT and had the midterm and final examinations of the following semester seven courses (i.e., the midterm and final examinations of Eng. 111, Eng. 112, etc.). The scores of students who failed the IEPPAT, transferred to other departments, or withdrew of the second semester of 1997/1998 were discarded.

3. Procedure

In part four, students were asked to write about either one of the two following prompts: "A friend in need is a friend indeed" or "Imagine that you had won 100,000 SR. How are you going to spend the money?" The students were allowed 30 minutes to write a minimum of a twelve-line paragraph. The composition was analytically scored by three different instructors of the intensive English program, and they were asked to follow the previously mentioned scoring schema. The time allotted for the midterm examinations of the seven courses was one hour whereas two hours were allocated for finals. The testing items used in these courses consisted of various techniques such as multiple-choice, fill-in-the-blank, and essay questions. It is noteworthy here that midterm and final examinations of the writing of the intensive English program were holistically marked.

D. Results and Discussion

The first step in the statistical analysis was to calculate the means (X), standard deviations (SD), and multiple correlation coefficients used as reliability indices. The first two statistics were calculated by getting first the average of the three different raters, then the regular statistical procedures were carried out. These statistics are displayed in Table 1.

Table 1. Means, standard deviations, and reliability indices of the six features of the analytic marking schema.

Feature/component	X	SD	R^2
Content	16.200	4.102	.77
Organization	10.582	1.423	.81
Vocabulary	7.545	1.152	.89
Grammar	12.873	3.067	.83
Punctuation	9.909	2.304	.79
Spelling	11.036	2.411	.83
Total	68.145	10.051	.84

As the table shows, the values of the standard deviations are low, which indicates that the deviations of students' scores from the means are low. The sample seems to be homogeneous. Furthermore, the value of the standard deviation of students' scores in the feature content could be justified on the basis that evaluating content is rather subjective. Yet, this subjectivity is not as high as the one that might be found if the holistic method is to be used. A closer look at the values of the means of students in the various features indicates that the performance of students in some features was better than their performance in other features. For example, if the mean of the feature content is transformed into a percentage, the mean percentage is 64.80, which is almost 4 points lower than students' total. By the same token, the mean percentage of vocabulary is 75.45 which again exceeds the total average by 7 points. This led to the question as to whether the means of some features were higher statistically than the means of other features. To be able to compare the six means, and since their maximum differs from one feature to another, scores in various features were transformed to have a maximum of 25, similar to the highest maximum (i.e., the maximum of the feature content). Hence, the scores of the feature organization were multiplied by 1.67 and the scores of the feature vocabulary were multiply by 2.5. The same thing was done for the rest of the scores of the other features. To compare the means of the six features, an analysis of variance (ANOVA) was computed as it appears in Table 2.

Table 2. Analysis of variance for the difference between the subjects' means in the six features.

Source of Variance	DF	SS	MS	F
Between subjects	5	390.431	78.087	6.126*
Within subjects	324	4130.161	12.747	
Total	329	4520.592		

* $P < .0001$

As the table indicates, F was significant at $p < .0001$. Hence, Scheffe's post hoc comparisons were calculated to detect the difference direction. The differences between the means and the values of Scheffe's F-test are given in Table 3.

Table 3. Values of Scheffe's F-test between the subjects' means in the six features of the analytic marking schema.

Comparisons	Means diff.	Scheffe F-test
Content vs. organization	-1.427	1.553
Content vs. vocabulary	-2.664	5.087**
Content vs. grammar	.109	.009
Content vs. punctuation	-.348	.087
Content vs. spelling	-2.231	3.568*
Organization vs. vocabulary	- 1. 192	1. 019
Organization vs. grammar	1.581	1. 791
Organization vs. punctuation	1.123	. 905
Organization vs. spelling	-. 759	. 413
Vocabulary vs. grammar	2. 773	5. 512*
Vocabulary vs. punctuation	2. 315	3. 844*
Vocabulary vs. spelling	. 433	. 134
Grammar vs.punctuation	-. 457	.150
Grammar vs. spelling	-2. 34	3. 925*
Punctuation vs. spelling	-1. 883	2. 541*

- $p < .05$
- $^{**} P < .01$

As appears from the table, the only significant differences were those between the means of vocabulary and content, spelling and content, vocabulary and grammar, vocabulary and punctuation, grammar and spelling, and punctuation and spelling. With such significant differences, the second null hypothesis of this study is thus rejected, and we may conclude that the performance of subjects in the various features of the analytic marking schema differs. A reconsideration of Table 3 reveals that students faced difficulty while dealing with the features' content, grammar, and punctuation. Their organization of their composition seems to be acceptable. In addition, it seems that they have no difficulty in selecting appropriate vocabulary and in spelling them correctly.

To investigate the intra-rater reliability, Cronbach alpha for each rater was calculated with each feature being a separate variable. The values of Cronbach alpha for the three raters were as follows: .88, .81, and .79. This indicates that students who scored high in one feature, for example organization, also score high in other features. Or when a rater rates a students high in one feature he also rates him high in other features. In addition, the values of Cronbach alpha reflect the suitability of the marking grid used. The distribution of the total score of composition over the six features appeared also to be optimal. Otherwise we would get low Cronbach alpha values. It is obvious that the six features work together to paint a good picture of the students' ability in the writing skill.

The inter-raters reliability was calculated using θ which was proposed by Winer [14] cited in Zughoul and Kambal [3,p. 99], and employed by Al-Fallay [7, p.7]. To obtain the inter-raters reliability, we get first □ value through the use of the following formula:

$$\theta = \frac{\text{MS between subject -MS within subjects}}{\text{(K) (MS within subjects}}$$

where MS stands for the mean of squares and K refers to the number of raters. The reliability index, r, is then calculated by employing the formula:

$$r\,3 = \frac{3\theta}{1+3\theta}$$

The value of r obtained was .80, and this value is not far-off the values of r obtained in similar studies.

To investigate the linear relationship among the various features of the marking schema, Pearson product-moment correlation coefficients were calculated as Table 4 indicates.

Table 4. Pearson product-moment correlation coefficients among the six features of the analytic marking schema*.

	1	2	3	4	5
1 Content					
2 Organization	.688				
3 Vocabulary	.483	.840			
4 Grammar	.613	.831	.829		
5 Punctuation	.510	.757	.783	.801	
6 Spelling	.374	.764	.847	.810	.860

* $p < .05$

With the exception of the correlation coefficients of the feature content and other features, all correlations are high. They ranged from .860, the correlation coefficient between spelling and punctuation, to .757, the coefficient of the correlation between organization and punctuation. Though the correlation coefficient of the feature content may seem low, all coefficients were statistically significant not only at $p < .05$ level of significance but at $p < .0001$. The low values of the coefficients of the feature content might be justified on the basis of the subjectivity of the rating of a composition's content. The highest coefficient of the feature content was the one between it and the feature organization. This could be ascribed to the fact that rating of organization is more subjective than rating of the other features such as spelling or punctuation. With the previous discussion in mind, the second null hypothesis may be rejected, and a conclusion that there are significant correlations among the six features of the analytic marking schema may be formed.

The correlation between the students' scores in the six features and their scores on corresponding teacher-made midterm and final examinations was calculated. In other words, the correlation between the subjects' scores in the feature grammar and their scores on grammar midterm and final examinations was calculated. By the same token, the subjects' scores in the feature content, organization, punctuation, and spelling were correlated with their scores on the writing midterm and final writing tests administered to the intensive English program students. Finally, the scores in the feature vocabulary were correlated with

the scores of subjects on the reading midterm and final examinations. The correlations appear in Table 5.

Table (5) Pearson product-moment correlation among the subjects' scores in the six features and their scores on corresponding teacher-made midterm and final examinations*

	Midterm reading	Final reading	Midterm writing	Final writing	Midterm grammar	Final grammar
1 Content			. 508	. 496		
2 Organization			. 594	. 611		
3 Vocabulary	. 598	. 701				
4 Grammar					. 704	. 654
5 Punctuation			. 617	. 597		
6 Spelling			. 694	. 733		

*$p < .05$

The teacher-made midterm and final writing tests were holistically marked. Yet, the correlations between these two tests and the features content, organization, punctuation, and spelling were statistically significant. As expected, the correlation coefficients between the features punctuation and spelling and the writing midterm and final were higher than those between the features content and organization and the same tests. The highest correlation obtained was between the feature vocabulary and the reading midterm and final examinations reflecting the strength of the linear relationship between the feature that the analytic marking was considering and the knowledge assessed by the reading tests. With the above mentioned discussion in mind, the third null hypothesis is thus rejected, and we may conclude the presence of a significant correlation between the scores of subjects in the features of the analytic marking schema and their scores on corresponding tests made by classroom teachers.

Since the scores of students on another parts of the IEPPAT were available, it was tempting to investigate if there was a relationship between the six features of the analytic scoring and similar parts geared to measure the same thing measured by these features. Hence, the feature grammar was correlated with the grammar part and the feature vocabulary was correlated with the part assessing knowledge of vocabulary. If we would correlate the other four features with the total of part four (i.e., subjects' scores in the writing part) we would get inflated indices because of the part-whole overlap. So, only the first two correlations were calculated. The correlation between the feature grammar and the grammar part of the IEPPAT was .704 which was significant at $p < .0001$. The correlation between the feature vocabulary and the vocabulary part of the IEPPAT was not as high as the former correlation; however, it was relatively high, $r = .634$, and it was significant at $p < .0001$.

To test the fourth hypothesis set by this study concerning factor loading of the six features on one or more factors, principle factor analysis was used. Considering the nature of the six features, they could be classified into two categories: (a) feature(s) related to proficiency in the foreign language or taught

skill features, and (a) feature(s) related to knowledge of the world or untaught skill features. In the first category, the features' organization, vocabulary, punctuation, and spelling may be listed. In the second category, the feature content can be listed since it is so hard to teach students how to make the content of their compositions more effective. However, it is easier to teach and master the features of the former category. This might explain the reason behind the presence of such topics as spelling, punctuation and grammar in writing textbooks and the absence of topics teaching the mastery of effective content. With such discussion in mind, a two-factor solution was proposed. The first factor is the general factor or the factor related to proficiency in the foreign language and the second factor is related to knowledge of the world. The loadings on these two factors are given in Table 6.

Table 6. Principle factor solution (without iteration) for the six features of the analytic marking schema of the study's subjects accounting for 77.94% of the variance.

Feature	Factor 1	Factor 2	h^2
1 Content	.1474	.9569	.9157
2 Organization	.7723	.2359	.5964
3 Vocabulary	.8585	-.2293	.7368
4 Grammar	.8492	.2462	.7211
5 Punctuation	.8262	-.0009	.6826
6 Spelling	.8380	.2909	.7022
Eigenvalue	3.4391	.9157	
Proportion of variance accounted for	.6878	.0916	

Notice that only significant loadings (i.e., ± .30) were used in the calculation of communality (h^2), eigenvalues, and the proportion of the variance accounted for. As the table shows, features related to proficiency in the foreign language loaded significantly on the general factor (Factor 1), which is responsible for mastering the foreign language. The feature content loaded significantly on Factor 2, which is related to knowledge of the world. However, with principle factor analysis, factors should be rotated in order to get the optimum loadings. Hence varimax rotation procedure was used. Table 7 displays the loadings on the two factors after rotation.

Table 7. Varimax rotated factor solution (without Iteration) for the six features of the analytic marking schema of the study's subjects accounting for 84.35% of the variance.

Feature	Factor 1	Factor 2	h^2
1 Content	.0173	.9680	.9370
2 Organization	.8335	.2376	.6947
3 Vocabulary	.8815	-.1117	.7770
4 Grammar	.9084	.2582	.8252
5 Punctuation	.8188	.1102	.6704
6 Spelling	.8814	-.2729	.7816
Eigenvalue	3.7489	.9370	
Proportion of variance accounted for	.7498	.0937	

After factor rotation, the picture is clear. The loadings of the five features on the general factor increased, and their loadings on the second factor decreased. By the same token, the loading of the feature content on Factor 2 increased, and its loading on Factor 1 decreased. In addition, the percentage of the variance that the general factor was capable of accounting for was 84.35 of the total variance.

In order to ensure that no further factors were left unextracted, residual correlation coefficients were calculated. Statistically speaking, if the values of all coefficients are less than ± .10, this means that all significant factors are extracted. Table 8 gives the residual correlation coefficients of the six features based on two factors.

With the aforementioned discussion in mind, it is obvious that the loadings of the features are logical. Features that can be taught loaded significantly on the general factor responsible for proficiency in the foreign language. The feature that cannot be taught loaded significantly on another factor that can be related to knowledge of the world. With the previous statistical analysis in kind, the fourth null hypothesis is thus retained. It is clear that the features of the analytic scoring schema load on different features. This, in fact, indicates that the selection of the features was successful.

Table 8. Residual correlations with the loadings on the two factors partialed out based on two factors.

	1	2	3	4	5
1 Content					
2 Organization	-.087				
3 Vocabulary	.026	-.0153			
4 Grammar	-.042	.0263	-.0184		
5 Punctuation	.028	-.0870	-.0423	-.0846	
6 Spelling	.058	-.0574	-.0780	-.0168	.0761

To investigate whether the features of the analytic scoring schema are capable of predicting current and future levels of performance of subjects on similar or corresponding tasks and skills, the multiple regression statistic was calculated for each feature.

1- The Feature Content

To investigate the relationship between the feature content and the current level of performance of subjects in writing tasks, multiple regression analysis was conducted. The feature content was the dependent variable and the teacher-made writing midterm and final examinations were the independent variables. The summary table of the analysis of variance (ANOVA) is given in Table 9A.

Table 9A. Multiple regression analysis of the feature content as a dependent variable and teacher-made midterm and final writing tests as independent variables.

Analysis of variance					
Effect	*df*	SS	MS	F	*P*
Regression	2	989.7420	494.8710	3.8776	<.0127
Residual	52	6635.9800	127.6150		
Total		7625.7220			

Variables	B	SE B	Beta	t	*P*
Writing midterm	.1886	.0652	.3248	2.8910	.0056
Writing final	.2153	.0714	.3387	3.0151	.0034

The relationship between the feature content and the future performance of students in a similar task was investigated by multiple regression analysis where the feature content was the dependent variable and the midterm and final examinations of Eng. 111 (Writing Skills) were the independent variables. Table 9B gives ANOVA summary table and regression weights.

Table 9B. Multiple regression analysis of the feature content as a dependent variable and midterm and final tests of Eng. 111 as independent variables.

Analysis of variance					
Effect	*Df*	SS	MS	F	*P*
Regression	2	933.8000	466.9000	4.7432	<.0028
Residual	52	5126.4746	98.5861		
Total		6060.2746			

Variables	B	SE B	Beta	T	*P*
Eng. 111 midterm	.3992	.0561	.1910	3.4046	.0016
Eng. 1111 final	.5099	.2220	.1724	3.0726	.0031

As Tables 9A and 9B indicate, there is a significant relationship between the feature content and tests of a corresponding nature. It seems possible to predict future and current levels of performance of subjects on tasks related to the feature content.

2- The Feature Organization

The relationship between the feature organization and the teacher-made midterm and final writing tests (current performance) was investigated. Results are summarized in Tables 10A and 10B.

Table 10A. Multiple regression analysis of the feature organization as a dependent variable and teacher-made midterm and final writing tests as independent variables.

Analysis of variance					
Effect	*Df*	SS	MS	F	*P*
Regression	2	410.2588	205.1294	15.6580	<.00001
Residual	52	681.2305	13.1006		
Total		1091.4893			

Variables	B	SE B	Beta	T	*P*
Writing midterm	.4306	.0400	.3358	8.3961	.000001
Writing final	.5923	.0536	.2046	5.1162	.000001

The significant relationship between the feature organization and the current and future performance on similar tasks is clear.

3- The Feature Vocabulary

Similar analysis was conducted with the feature vocabulary as dependent variable and the vocabulary section of the IEPPAT as independent variable. Table 11A gives the result of the linear regression analysis that was conducted.

Table 10B. Multiple regression analysis of the feature organization as a dependent variable and Eng. 111 midterm and final tests as independent variables.

Analysis of variance					
Effect	*Df*	SS	MS	F	*P*
Regression	2	150.7170	78.5850	4.3628	<.0072
Residual	52	936.6480	18.0120		
Total		1093.8180			

Variables	B	SE B	Beta	t	*P*
Writing midterm	.5069	.0523	.1460	2.7933	.0073
Writing final	.2351	.0255	.2554	4.8863	.00001

Table 11A. Linear regression analysis of the feature vocabulary as a dependent variable and the vocabulary section of the IEPPAT as an independent variable.

Analysis of variance					
Effect	*Df*	SS	MS	F	*P*
Regression	1	9.2012	9.2012	7.8180	<.0002
Residual	5	62.4351	1.1780		
Total		71.6364			

Variables	B	SE B	Beta	T	*P*
Vocabulary section of the IEPPAT	.0760	.0272	.3584	2.7948	.0072

The relationship between the feature vocabulary and Eng. 120 midterm and final examinations was investigated by multiple regression analysis as displayed in Table 11B.

4- The Feature Grammar

The feature grammar was also treated as a dependent variable in linear regression analysis where the grammar section of the IEPPAT was treated as the independent variable, and in the multiple regression analysis where Eng. 116 midterm and final examinations were treated as independent variables. Results of the analysis are displayed in Tables 12A and 12B.

Table 11B. Multiple regression analysis of the feature vocabulary as a dependent variable and Eng. 120 midterm and final tests as independent variables.

Analysis of variance					
Effect	*Df*	SS	MS	F	*P*
Regression	2	64.5832	32.2916	4.0380	<.0093
Residual	52	415.8403	7.9969		
Total		484.4235			

Variables	B	SE B	Beta	t	*P*
Eng. 120 midterm	.1156	.0342	.4431	3.3801	.0015
Eng. 120 final	.2127	.0352	.7930	6.0492	.000001

Table 12A. Linear regression analysis of the feature grammar as a dependent variable and the grammar section of the IEPPAT as an independent variable.

Analysis of variance					
Effect	*Df*	SS	MS	F	*P*
Regression	1	233.5331	233.5331	45.0777	<.00000
Residual	53	274.5760	5.1807		
Total		508.1091			

Variables	B	SE B	Beta	t	*P*
Grammar section	.4035	.0601	.6779	6.7139	
of the IEPPAT	.000001				

Table 12B. Multiple regression analysis of the feature grammar as a dependent variable and Eng. 116 midterm and final tests as independent variables.

Analysis of variance					
Effect	*Df*	SS	MS	F	*P*
Regression	2	175.0098	87.5049	11.7238	<.000017
Residual	52	388.1228	7.4639		
Total		563.1326			

Variables	B	SE B	Beta	T	*p*
Eng. 116 midterm	.5219	.1013	.3249	5.1520	.000001
Eng. 116 final	.4850	.0665	.4599	7.2932	.000001

Similar analysis was conducted with the features punctuation and spelling as dependent variables and the results of the analyses are given in Appendices 1 and 2.

The fifth null hypothesis was formulated to test the relationship between the features of the analytic marking schema and variables reflecting the current and future levels of students' proficiency in skills and languages element assessed by tests similar to the tasks assessed by the features. With the results of the statistical analyses in Tables 9A-12B and in Appendices 1 and 2 in mind, the hypothesis is thus rejected. The features are capable of predicting students' performance in corresponding tests; and the features of the analytic marking schema and the analytic approach seem to be an invaluable tool for pedagogical purposes, not only in assessing writing ability but in evaluating other related skills and language elements.

E. Summary and Conclusions

This study aimed at a close consideration of the analytic approach to marking composition. Most of the time, the use of the composition scores is limited to inform students, usually by the means of grades or numbers, about their ability in the writing skill or to determine whether students are ready to advance to higher stages in the learning process. The use of test scores should be broadened. Test scores should mean more than mere indices of students, language achievement or language proficiency. The close consideration of the analytic approach in composition scoring turned out to be invaluable. The six features of the analytic marking schema were selected after considering similar studies reported in the literature. The feature content was allotted 25% of the total score of composition because of the following two reasons: first, the subjects' level of proficiency in English as a foreign language was relatively low. Second, content is usually the feature to which much of the composition's score is usually assigned in holistic marking. The feature organization was allocated 15% of the total score because it reflects the ability of the writer in arranging the presentation of his idea. By the same token, punctuation and spelling were allotted 15% of the points because they are two essential components that raters consider when marking compositions. The features grammar and vocabulary were allotted 20% and 10% respectively.

The first finding of this study was that the analytic approach could be used to investigate the students' points of weakness and strength. The means of subjects in the six features differed significantly. The highest mean was the mean in the feature vocabulary, followed by the means in the features spelling, organization, punctuation, content, and grammar respectively. This indicates that the subjects did not face difficulty in selecting and spelling the used vocabulary. They seemed also to have no difficulties in organizing their compositions. However, their means in

the features punctuation, content, and grammar were relatively low. Their inability to correctly punctuate their paragraphs is expected at this low level of proficiency in the foreign language. This justification can also be extended to include the feature grammar. However, their low mean in the feature content needs further investigation. One reasonable speculation could be that the efforts which subjects took to produce an organized misspelling-free composition came at the expanse of its content. In general, such findings may guide the instructors to use certain remedial exercises and applications and may also remind material developers of the needs of students at this stage of foreign language proficiency.

The six features could be classified into types: features related to skills in the foreign language and those related to the general knowledge of the world. The feature content is the only example of the latter type. The correlations among the six features were relatively high. However, the lowest correlation coefficients were the ones between the feature content and the rest of the features. This was confirmed by the principle factor analysis with varimax rotation. The feature content loaded significantly on a second factor whereas the other features loaded significantly on a general factor related to the subjects' level of foreign language proficiency. Residual correlation coefficients indicate that no further factors are left unextracted. It seems that some writing instructors teach and emphasize the features related to proficiency in the foreign language. But they almost ignore the feature content which they either do not believe to be an important component of the writing skill, or do not know how to teach their students how to write effective compositions.

The relationship between the six features and other related courses were examined by conducting multiple and linear regression analyses among the scores in these features and the scores in these courses and on similar tests. Statistically significant relationship were found between the six features and similar courses and corresponding tests. It seems possible to use the scores in these features to predict current and future levels of performance in foreign language skills and language elements. The students' command of the grammar of English as a foreign language can be predicted by considering their scores in the feature grammar. Such a finding reflects the appropriateness of these features as components of a writing marking schema. In addition, the results of the analytic marking could also be utilized for diagnostic purposes, as mentioned earlier or for screening purposes of newly admitted students to programs of English as a foreign language.

Appendix 1.

Table 1A. Multiple regression analysis of the feature punctuation as a dependent variable and teacher-made midterm and final writing tests as independent variables.

Analysis of variance

Effect	*Df*	SS	MS	F	*P*
Regression	2	74.7697	37.3848	9.1795	<.000038
Residual	52	211.7758	4.0726		
Total		286.5454			

Variables	B	SE B	Beta	t	*P*
Writing midterm	.1486	.0561	.2708	2.6505	.0091
Writing final	.0999	.0275	.3712	3.6329	.00064

Table 1B. Multiple regression analysis of the feature punctuation as a dependent variable and Eng. 111 midterm and final tests as independent variables.

Analysis of variance

Effect	*df*	SS	MS	F	*P*
Regression	2	44.8589	22.4295	4.8258	<.0019
Residual	52	241.6865	4.6478		
Total		286.5454			

Variables	B	SE B	Beta	t	*P*
Writing midterm	.4079	.0839	.3343	4.8567	.00001
Writing final	.1839	.0410	.3891	4.4867	.00001

Appendix 2.

Table 2A. Multiple regression analysis of the feature spelling as a dependent variable and teacher-made midterm and final writing tests as independent variables.

Analysis of variance

Effect	*df*	SS	MS	F	*P*
Regression	2	120.3209	60.1605	16.1583	<.00001
Residual	52	193.6064	3.7232		
Total		313.9273			

Variables	B	SE B	Beta	t	*P*
Writing midterm	.2327	.0873	.2091	2.6655	.0092
Writing final	.1340	.0405	.2673	3.4069	.00137

Table 2B. Multiple regression analysis of the feature spelling as a dependent variable and Eng. 117 midterm and final tests as independent variables.

Analysis of variance

Effect	*df*	SS	MS	F	*P*
Regression	2	226.7338	113.3669	5.0449	<.0007
Residual	5	1168.5181	22.4715		
Total		1395.2519]			

Variables	B	SE B	Beta	t	*P*
Writing midterm	.2089	.0424	.5990	4.9312	.00001
Writing final	.2449	.0357	.8317	6.8473	.000001

References

[1] The Educational Testing Service. *TOEFL Bulletin of Information*. Princeton, N.J.: ETS. 1995.

[2] Hadley, A. *Teaching Language in Context*. Boston, MA: Heinle & Heinle Publishers,1993.

[3] Zugoul, M., and M. Kambal. "Objective Evaluation of EFL Composition." *International Review of Applied Linguistics*, 21 (1983), 87-103.

[4] Cooper, C. "Holistic Evaluation of Writing. " In *Evaluating Writing: Describing, Measuring, Judging,* edited by C. Cooper and L. Idell (Urbana, IL: National Council of Teachers of English, 1977), 3-31.

[5] Perkins, K. "On the Use of Composition Scoring Techniques, Objective Measures, and Objective Tests to Evaluate ESL Writing Ability. " *TESOL Quarterly*, 17 (1983), 651-71.

[6] Heaton, B. *Writing English Language Tests*. London: Longman,1991.

[7] Al-Fallay, I. "Validating a Multiple-Choice Cloze Test to Assess the Proficiency of EFL Learners' Writing Skill." *The Arab Journal of Humanities*, 65 (1998), 273-306.

[8] Cleary, C. "Testing Lower Intermediate Writing: A Comparison of Two Scoring Methods." *British Journal of Language Teaching*, 26 (1988), 75-80.

[9] Al-Gusair, F. "The Teaching and Evaluation of English Composition Writing at Female Public Secondary Schools in Riyadh." Unpublished master's thesis, King Saud University, Riyadh, Saudi Arabia, 1997.

[10] Follamn, J., and J. Anderson. "An Investigation of the Reliability of Five Procedures for Grading English Theme." *Research in the Teaching of English*, 1 (1967), 190-200.

[11] Kaczmarek, C. "Scoring and Rating Essay Talks." In *Research in Language Testing*, edited by J. Oller and K. Perkins. Rowley, MA: Newbury House Publishers, 1980, 151-59.

[12] Gilfert, S., and K. Harada. "Two Composition Scoring Methods: The Analytic vs. Holistic Method." *Bulletin of Faculty of Foreign Languages*, 1 (1992), 17-22.

[13] Nyberg, V., and A. Nyberg. "Alberta Essay Scale Models." Alberta University. ERIC Document No. ED220483, 1982 .

[14] Winer, J. *Statisical Principles in Expermental Design*. New York: McGraw- Hill, 1971.

تقويم طريقة التصحيح التحليلي: تطوير واستخدام نظام التصحيح التحليلي

إبراهيم صالح الفلاي

أستاذ مشارك، قسم اللغة الإنجليزية ، كلية الآداب
جامعة الملك سعود، الرياض، المملكة العربية السعودية

(قدم للنشر في ١٤١٩/٦/٨ هـ، قبل للنشر في ١٤٢٠/٢/١٠ هـ)

ملخص البحث. تهدف هذه الدراسة إلى تناول بالبحث الدقيق الطريقة التحليلية في تصحيح الإنشاء وجدواها كوسيلة تعليمية. وقد تم جمع بيانات هذه الدراسة في ثلاثة مصادر(١) النتائج التي حصل عليها ٥٥ طالبا سعوديا يدرسون في برنامج اللغة الإنجليزية بكلية الآداب – جامعة الملك سعود بالرياض بالمملكة العربية السعودية، وذلك في الأجزاء الخاصة بالكتابة والقواعد والمفردات من اختبار الكفاءة/التحصيل في اللغة الإنجليزية الخاص ببرنامج الإنجليزية المكثف، (٢) نتائج نفس الطلاب في اختبارات الكتابة والقواعة والقواعد التي قام بإعدادها مدرسو برنامج الإنجليزية المكثف كاختبارات لمنتصف الفصل الدراسي ونهايته، (٣) نتائج نفس الطلاب في المواد التي تدرس في القسم خلال الفصل الدراسي التالي لبرنامج الإنجليزية المكثف.

وقد توصلت الدراسة إلى أن طريقة التصحسح التحليلي تتميز بمعاملات عالية من الصدق بين وضمن المقيمين. ووحدات الدراسة كذلك أنه من الملائم توزيع الدرجات على أساس ستة مكونات أو عناصر هي: المحتوى والتنظيم والمفردات والقواعد واستخدام علامات الترقيم والتهجئة، وأن الدرجات المخصصة لكل عنصر كانت أيضا مناسبة.

وبينت الدراسة أن من الأسهل بالنسبة للطلاب التعامل مع عناصر المفـــردات والتهجئـــة والتنظيم مقارنة بعناصر استخدام علامات الترقيم والمحتوى والقواعد. وكشفت الدراســـة كذلك عن وجود علاقات قوية بين العناصر المختلفة. وقد تم تحميل هذه العنـــاصر علـــى عاملين مختلفين: الول وثيق الصلة بالكفاءة في اللغة الإنجليزية؛ أما الثـــانيْ فـــهو أكـــثر اتصالا بالمعلومات العامة عن العالم. وثبت كذلك أن من الممكن التنبـــؤ بـــأداء الطـــلاب الحالي والمستقبلي على أساس اختبارات تقيس هذه العناصر، ومن بيـــن توصيـــات هـــذه الدراسة أن استخدام طريقة التصحيح التحليلي أمر في غاية الأهمية، وذلك لأنه باتباع هـذه الطريقة يمكن الاستفادة من النتائج على نطاق أوسع. ومن بين التوصيات أيضا أنـــه فـــي إمكان مدرسي الكتابة مساعدة طلابهم على تحسين أدائهم في كتابة مواضيع إنشـــاء ذات محتوى فعال. وفي النهاية تقترح الدراسة أن يفحص مدرسو الكتابة نتائج طلابهم في هـذه العناصر الستة للكشف عن أوجه القوة والضعف، وعليه يمكن تصميم تدريبـــات علاجيـــة لأوجه الضعف، وأن يأخذ مصممو المناهج في الاعتبار هذه الأمور عند تصميـــم منـــاهج الكتابة، خصوصا في البرامج المكثفة للإنجليزية كلغة أجنبية.